CSE 258

Web Mining and Recommender Systems

Midterm recap

Midterm on Wednesday!

- 6:40 pm – 7:40 pm
- Closed book – but I'll provide a similar level of basic info as in the last page of previous midterms

CSE 258

Web Mining and Recommender Systems

Week 1 recap

Supervised versus unsupervised learning

Learning approaches attempt to **model data** in order to solve a problem

Unsupervised learning approaches find patterns/relationships/structure in data, but **are not** optimized to solve a particular predictive task

- E.g. PCA, community detection

Supervised learning aims to directly model the relationship between input and output variables, so that the output variables can be predicted accurately given the input

- E.g. linear regression, logistic regression

Linear regression

Linear regression assumes a predictor of the form

$$X\theta = y$$

matrix of features (data)

unknowns (which features are relevant)

vector of outputs (labels)

(or $Ax = b$ if you prefer)

Regression diagnostics

Mean-squared error (MSE)

$$\frac{1}{N}\|y - X\theta\|_2^2$$

$$= \frac{1}{N}\sum_{i=1}^{N}(y_i - X_i \cdot \theta)^2$$

Representing the month as a feature

How would you build a feature to represent the **month?**

Representing the month as a feature

Occam's razor

"Among competing hypotheses, the one with the fewest assumptions should be selected"

Regularization

Regularization is the process of penalizing model complexity during training

$$\arg\min_\theta = \frac{1}{N}\|y - X\theta\|_2^2 + \lambda\|\theta\|_2^2$$

How much should we trade-off accuracy versus complexity?

Model selection

A **validation set** is constructed to "tune" the model's parameters

- Training set: used to **optimize the model's parameters**
- Test set: used to report how well we expect the model to perform on **unseen data**
- Validation set: used to **tune** any model parameters that are not directly optimized

Regularization

Model selection

A few "theorems" about training, validation, and test sets

- The training error **increases** as lambda **increases**
- The validation and test error are at least as large as the training error (assuming infinitely large random partitions)
- The validation/test error will usually have a "sweet spot" between under- and over-fitting

CSE 258 – Lecture 10

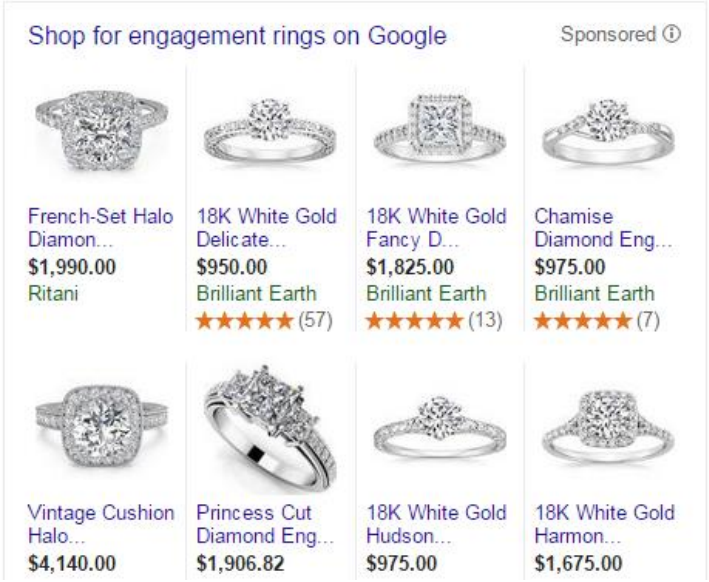
Web Mining and Recommender Systems

Week 2

Classification

Will I **purchase** this product?

(yes)



Will I **click on** this ad?

(no)

Classification

What animal appears in this image?

(mandarin duck)

Classification

What are the **categories** of the item being described?

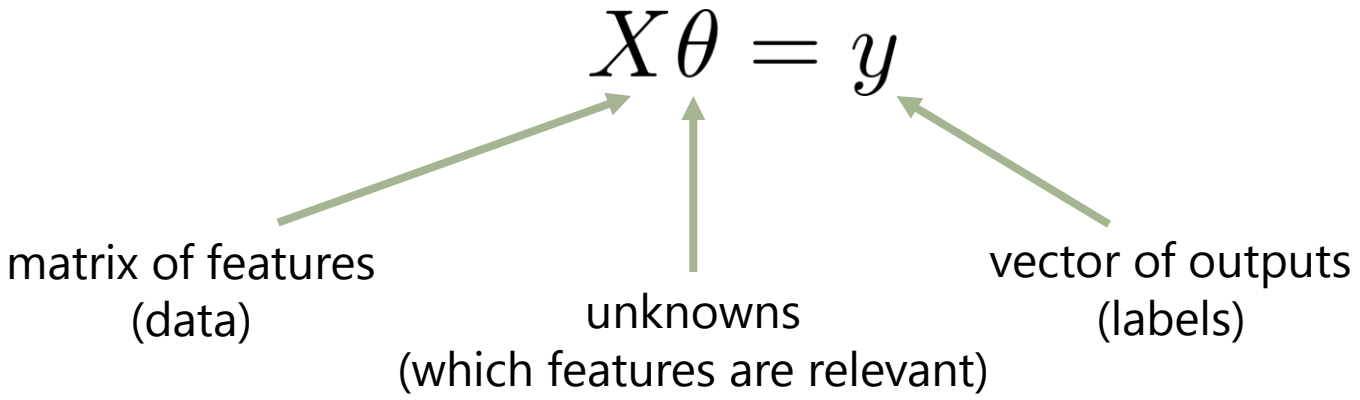
(book, fiction, philosophical fiction)

From Booklist

Houellebecq's deeply philosophical novel is about an alienated young man searching for happiness in the computer age. Bored with the world and too weary to try to adapt to the foibles of friends and coworkers, he retreats into himself, descending into depression while attempting to analyze the passions of the people around him. Houellebecq uses his nameless narrator as a vehicle for extended exploration into the meanings and manifestations of love and desire in human interactions. Ironically, as the narrator attempts to define love in increasingly abstract terms, he becomes less and less capable of experiencing that which he is so desperate to understand. Intelligent and well written, the short novel is a thought-provoking inspection of a generation's confusion about all things sexual. Houellebecq captures precisely the cynical disillusionment of disaffected youth. *Bonnie Johnston --This text refers to an out of print or unavailable edition of this title.*

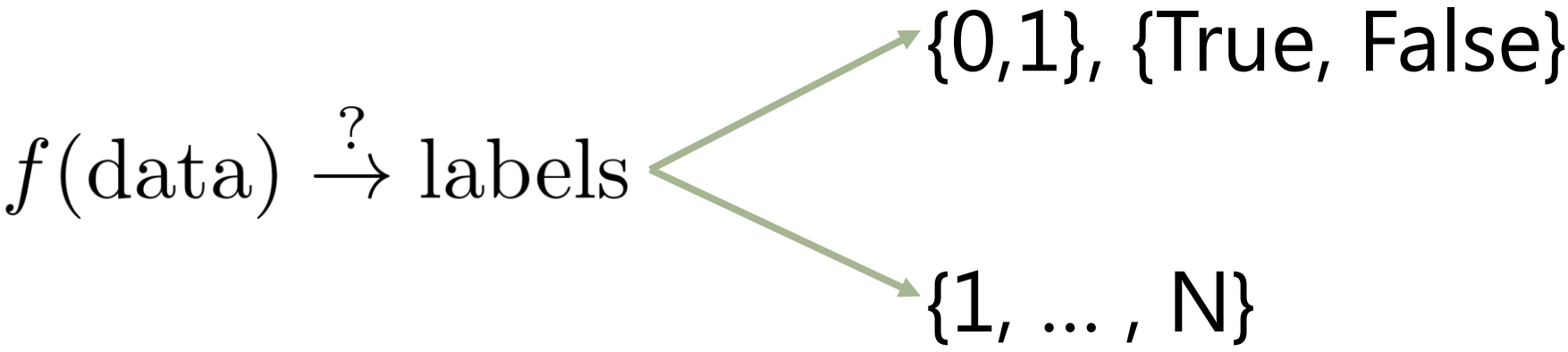
Linear regression

Linear regression assumes a predictor of the form



Regression vs. classification

But how can we predict **binary** or **categorical** variables?



(linear) classification

We'll attempt to build **classifiers** that make decisions according to rules of the form

$$y_i = \begin{cases} 1 & \text{if } X_i \cdot \theta > 0 \\ 0 & \text{otherwise} \end{cases}$$

In week 2

1. Naïve Bayes

Assumes an **independence** relationship between the features and the class label and "learns" a simple model by counting

2. Logistic regression

Adapts the **regression** approaches we saw last week to binary problems

3. Support Vector Machines

Learns to classify items by finding a hyperplane that separates them

Naïve Bayes (2 slide summary)

$$(feature_i \perp\!\!\!\perp feature_j | label)$$

$$p(feature_i, feature_j | label)$$

$$=$$

$$p(feature_i | label) p(feature_j | label)$$

Naïve Bayes (2 slide summary)

Double-counting: naïve Bayes vs Logistic Regression

Q: What would happen if we trained two regressors, and attempted to "naively" combine their parameters?

$$\text{no. of pages} = \alpha + \beta_1 \cdot \delta(\text{mentions wizards})$$

$$\text{no. of pages} = \alpha + \beta_2 \cdot \delta(\text{mentions witches})$$

$$\text{no. of pages} = \alpha + \beta_1 \cdot \delta(\text{mentions wizards}) + \beta_2 \cdot \delta(\text{mentions witches})$$

Logistic regression

sigmoid function: $\sigma(t) = \frac{1}{1+e^{-t}}$

Logistic regression

Training:

$X_i \cdot \theta$ should be maximized when y_i is positive and minimized when y_i is negative

$$\arg\max_\theta \prod_i \delta(y_i = 1) p_\theta(y_i | X_i) + \delta(y_i = 0)(1 - p_\theta(y_i | X_i))$$

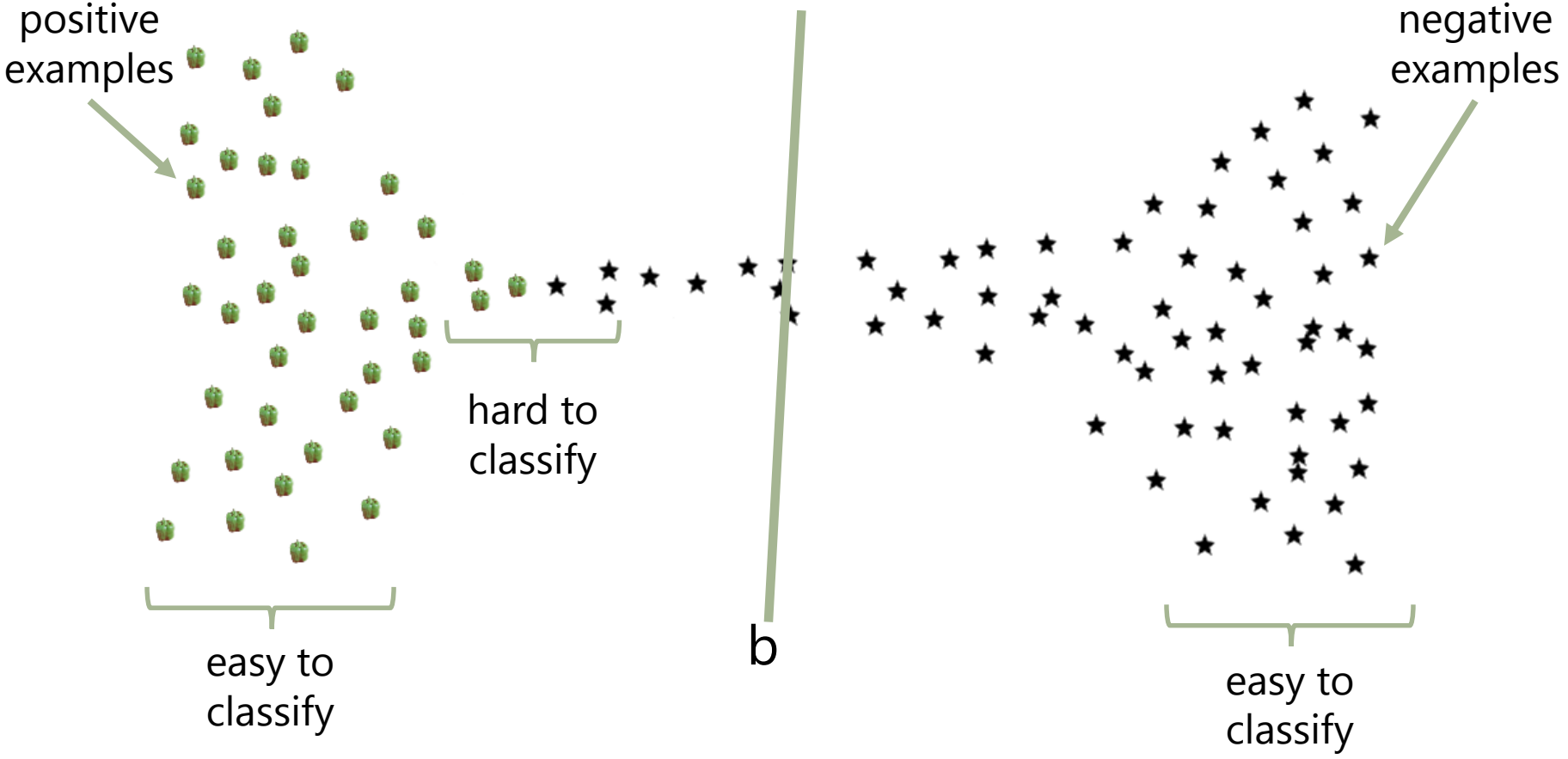
$\delta(\arg)$ = 1 if the argument is true, = 0 otherwise

Logistic regression

$$\arg\max_\theta \prod_i \delta(y_i = 1) p_\theta(y_i|X_i) + \delta(y_i = 0)(1 - p_\theta(y_i|X_i))$$

Logistic regression

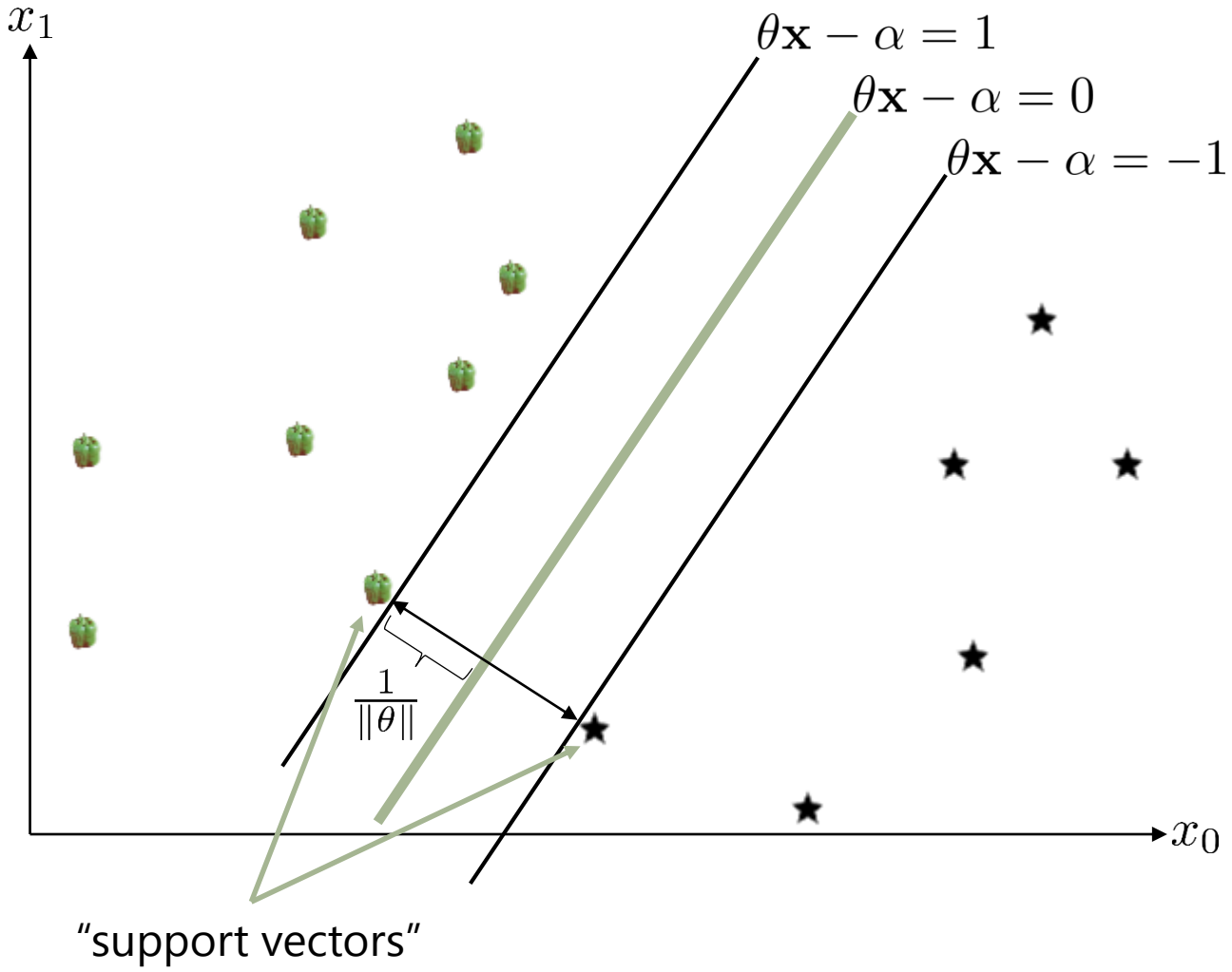
Q: Where would a logistic regressor place the decision boundary for these features?



Logistic regression

- Logistic regressors don't optimize the number of "mistakes"
- No special attention is paid to the "difficult" instances – every instance influences the model
- But "easy" instances can affect the model (and in a bad way!)
- How can we develop a classifier that optimizes the number of mislabeled examples?

Support Vector Machines



$$\arg\min_{\theta,\alpha} \frac{1}{2}\|\theta\|_2^2$$

such that

$$\forall_i y_i(\theta \cdot X_i - \alpha) \geq 1$$

Summary

The classifiers we've seen in Week 2 all attempt to make decisions by associating weights (theta) with features (x) and classifying according to

$$y_i = \begin{cases} 1 & \text{if } X_i \cdot \theta > 0 \\ 0 & \text{otherwise} \end{cases}$$

Summary

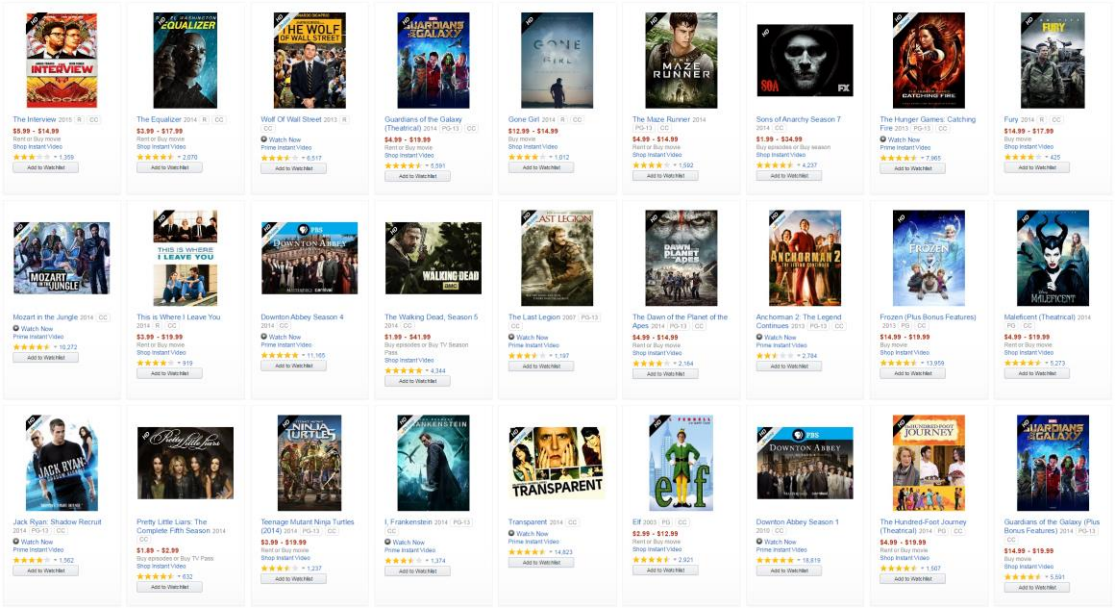
- **Naïve Bayes**
  - Probabilistic model (fits $p(label|data)$)
  - Makes a conditional independence assumption of the form $(feature_i \perp\!\!\!\perp feature_j|label)$ allowing us to define the model by computing $p(feature_i|label)$ for each feature
  - Simple to compute just by counting
- **Logistic Regression**
  - Fixes the "double counting" problem present in naïve Bayes
- **SVMs**
  - Non-probabilistic: optimizes the classification error rather than the likelihood

Which classifier is best?

1. When data are highly imbalanced

If there are far fewer positive examples than negative examples we may want to assign additional weight to negative instances (or vice versa)

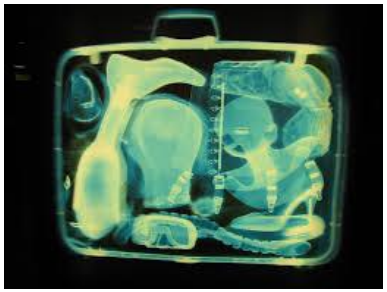
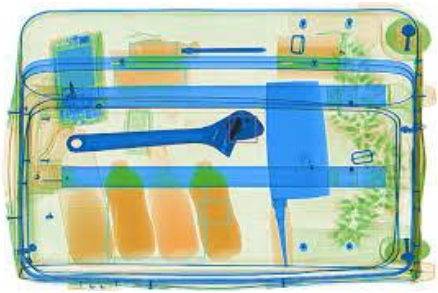
e.g. will I purchase a product? If I purchase 0.00001% of products, then a classifier which just predicts “no” everywhere is 99.99999% accurate, but not very useful



Which classifier is best?

2. When mistakes are more costly in one direction

False positives are nuisances but false negatives are disastrous (or vice versa)

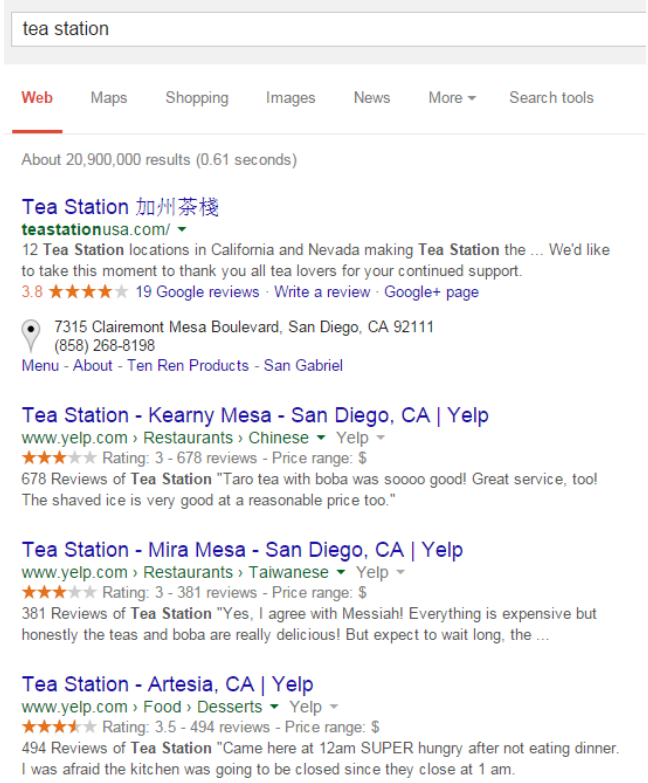



e.g. which of these bags contains a weapon?

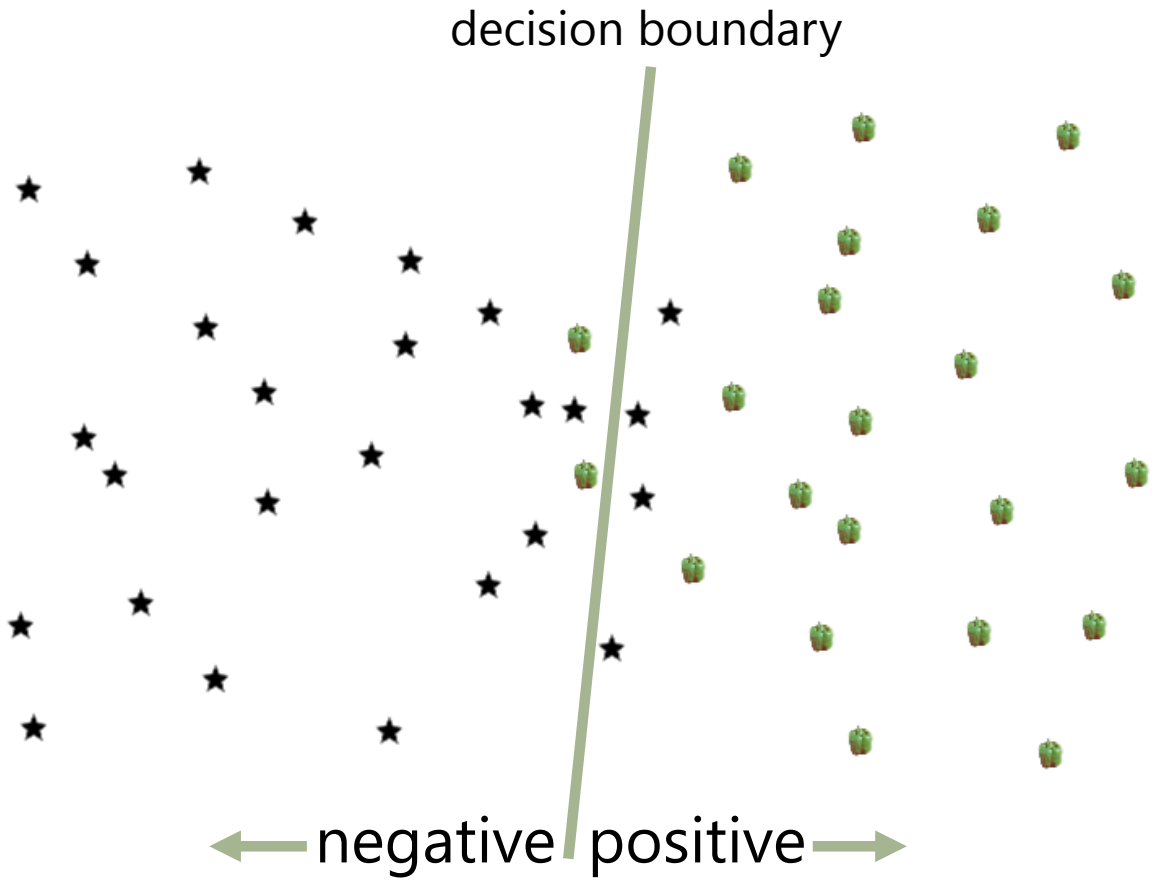
Which classifier is best?

3. When we only care about the "most confident" predictions

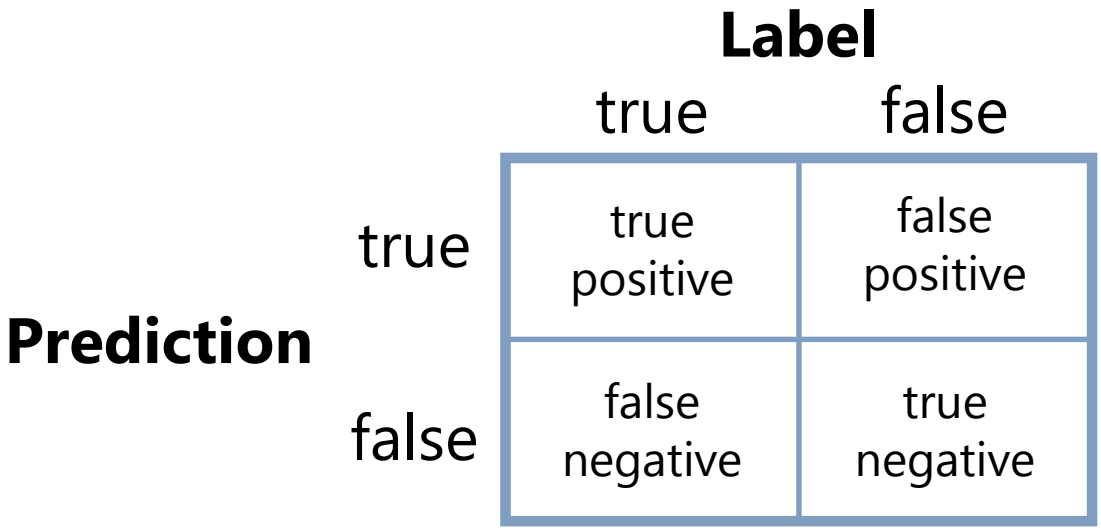
e.g. does a relevant result appear among the first page of results?



Evaluating classifiers



Evaluating classifiers



Classification accuracy = correct predictions / #predictions
= (TP + TN) / (TP + TN + FP + FN)

Error rate = incorrect predictions / #predictions
= (FP + FN) / (TP + TN + FP + FN)

Week 2

- Linear classification – know what the different classifiers are and when you should use each of them. What are the advantages/disadvantages of each
- Know how to evaluate classifiers – what should you do when you care more about false positives than false negatives etc.

CSE 258

Web Mining and Recommender Systems

Week 3

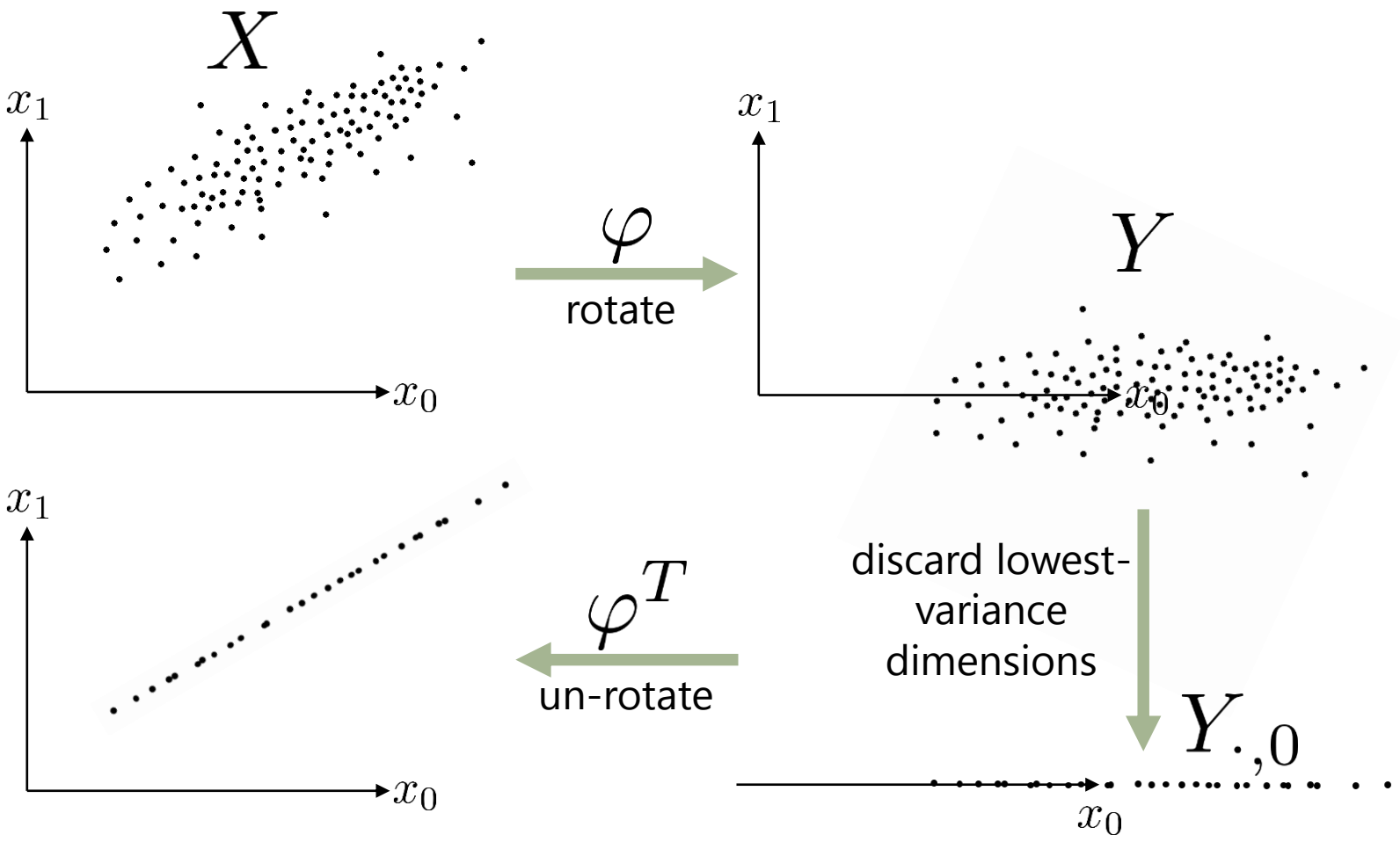
Why dimensionality reduction?

Goal: take **high-dimensional** data, and describe it compactly using a small number of dimensions

Assumption: Data lies (approximately) on some l**ow-dimensional** manifold

(a few dimensions of opinions, a small number of topics, or a small number of communities)

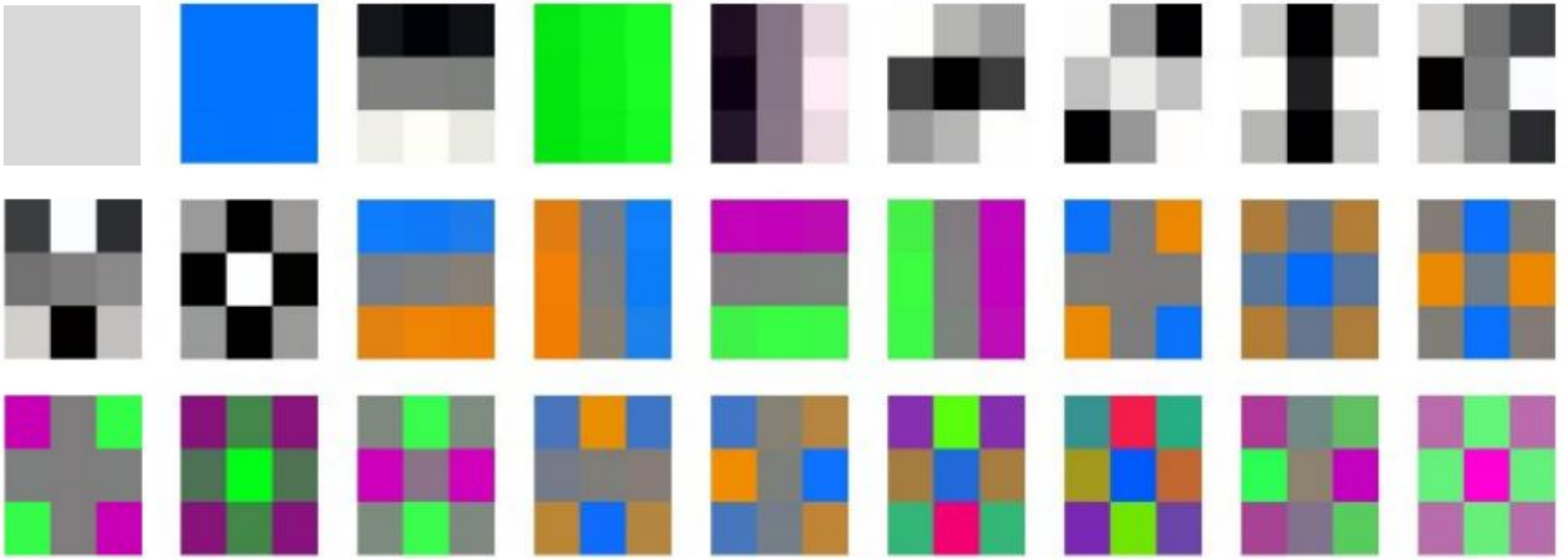
Principal Component Analysis



Principal Component Analysis

Construct such vectors from 100,000 patches from real images and run PCA:

Color:



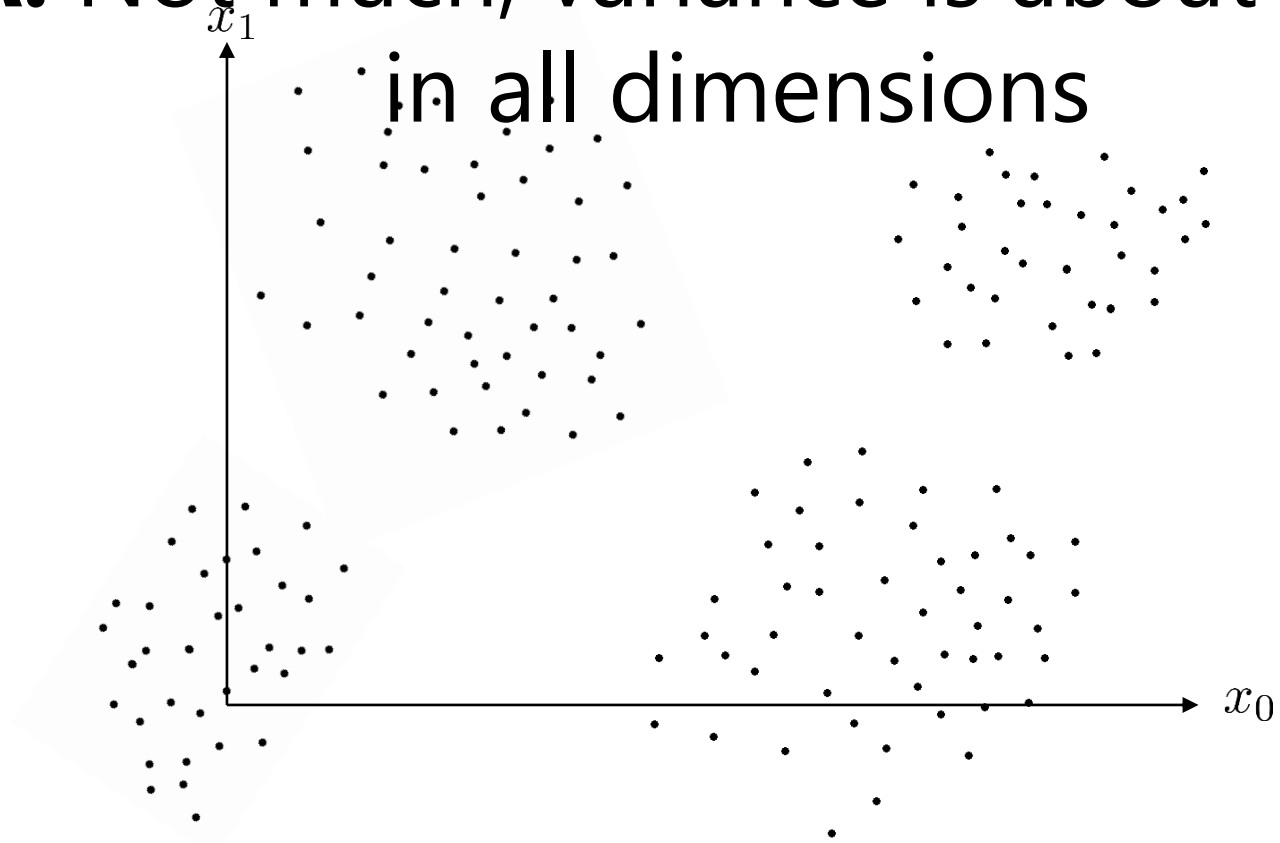
Principal Component Analysis

- We want to find a low-dimensional representation that best compresses or "summarizes" our data
- To do this we'd like to keep the dimensions with the highest variance (we proved this), and discard dimensions with lower variance. Essentially we'd like to capture the aspects of the data that are "hardest" to predict, while discard the parts that are "easy" to predict
- This can be done by taking the eigenvectors of the covariance matrix (we didn't prove this, but it's right there in the slides)

Clustering

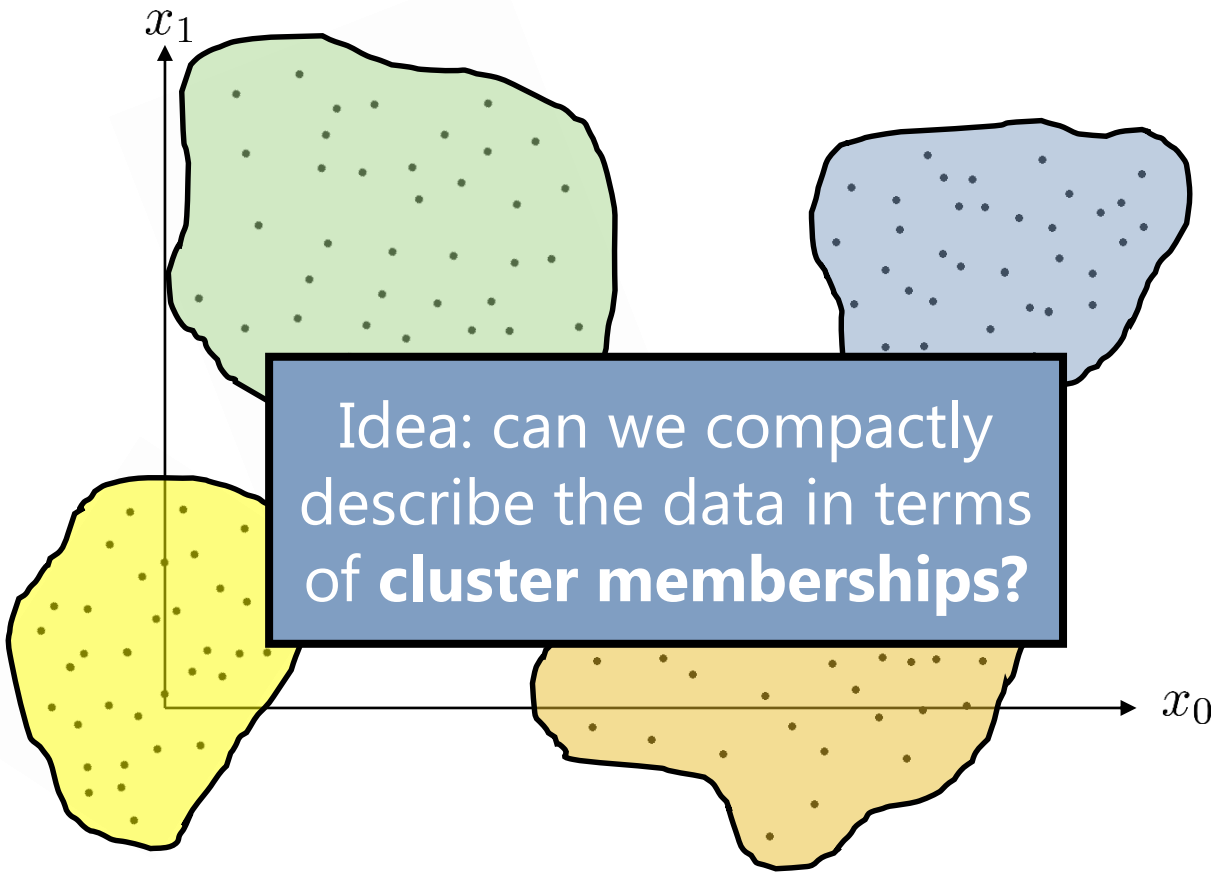
Q: What would PCA do with this data?

A: Not much, variance is about equal in all dimensions



Clustering

But: The data are highly **clustered**

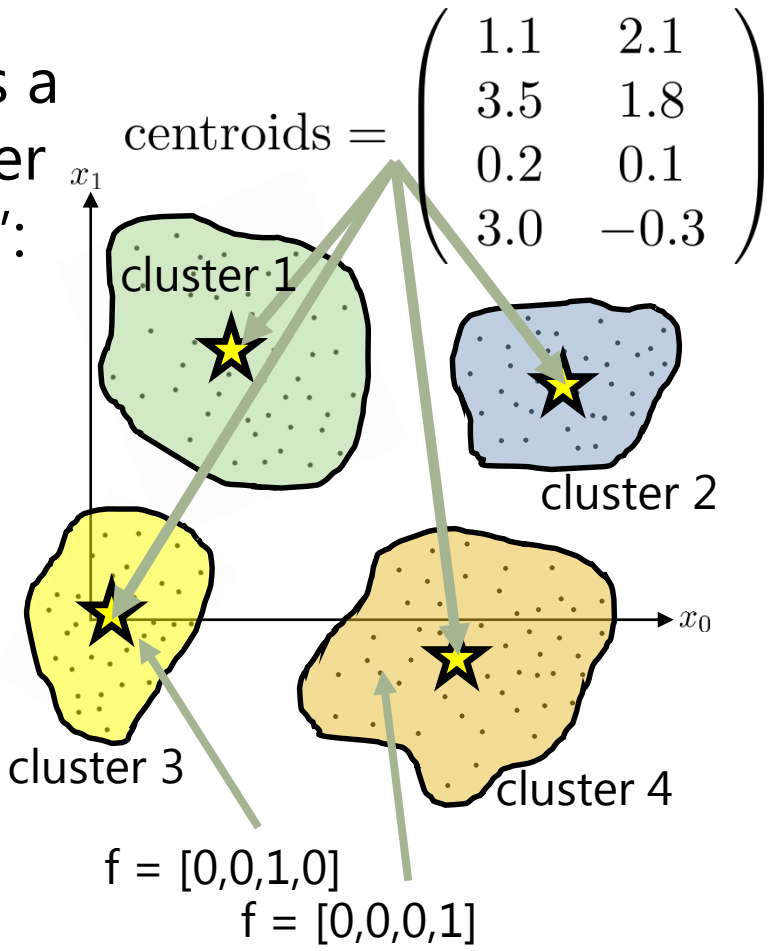


K-means Clustering

1. Input is still a matrix of features:

$$X = \begin{pmatrix} 5 & 3 & \cdots & 1 \\ 4 & 2 & & 1 \\ 3 & 1 & & 3 \\ 2 & 2 & & 4 \\ 1 & 5 & & 2 \\ \vdots & & \ddots & \vdots \\ 1 & 2 & \cdots & 1 \end{pmatrix}$$

2. Output is a list of cluster "centroids":



3. From this we can describe each point in X by its cluster membership:

$$Y = (1, 2, 4, 3, 4, 2, 4, 2, 2, 3, 3, 2, 1, 1, 3, \ldots, 2)$$

K-means Clustering

Greedy algorithm:

```
1. Initialize C (e.g. at random)
2. Do
3.       Assign each X_i to its nearest centroid
4.       Update each centroid to be the mean
         of points assigned to it
5. While (assignments change between iterations)
```

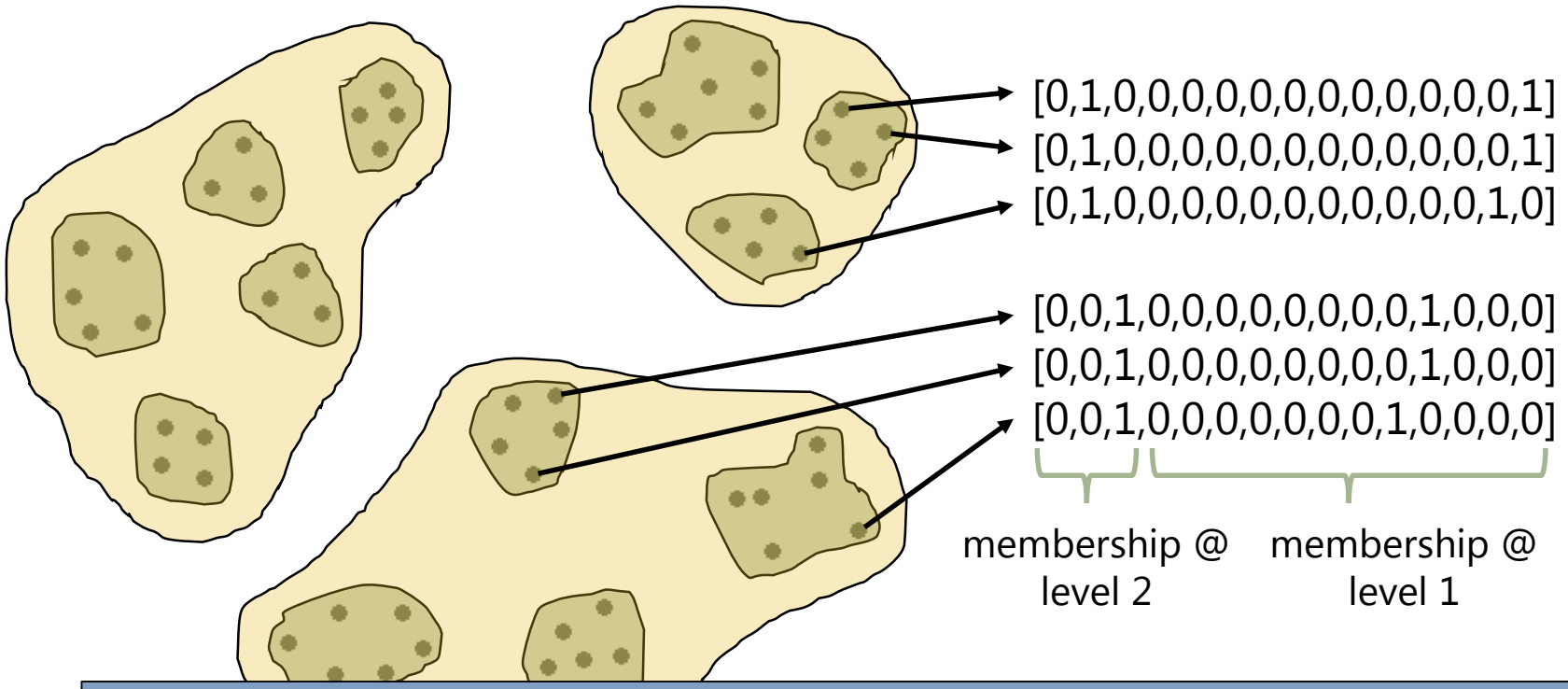
$$y_i = \arg\min_k \|X_i - C_k\|_2^2$$

$$C_k = \frac{\sum_i \delta(y_i=k) X_i}{\sum_i \delta(y_i=k)}$$

(also: reinitialize clusters at random should they become empty)

Hierarchical clustering

Q: What if our clusters are **hierarchical?**



A: We'd like a representation that encodes that points have **some features** in common but not others

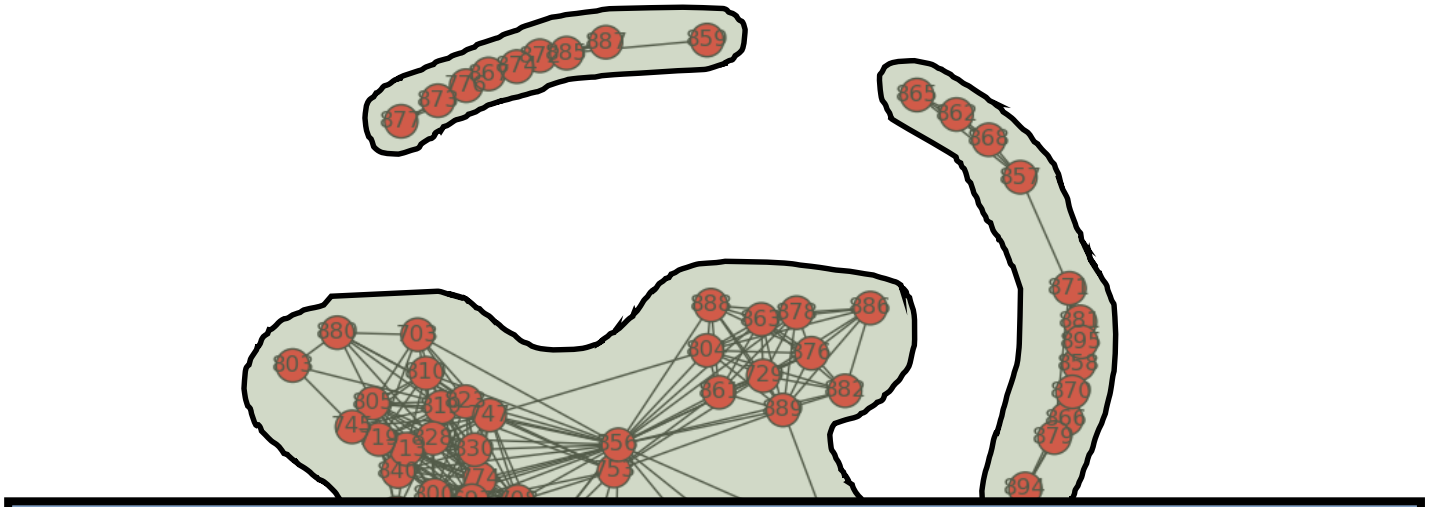
Hierarchical clustering

Hierarchical (agglomerative) clustering works by gradually fusing clusters whose points are closest together

```
Assign every point to its own cluster:
Clusters = [[1],[2],[3],[4],[5],[6],…,[N]]
While len(Clusters) > 1:
        Compute the center of each cluster
        Combine the two clusters with the nearest centers
```

1. Connected components

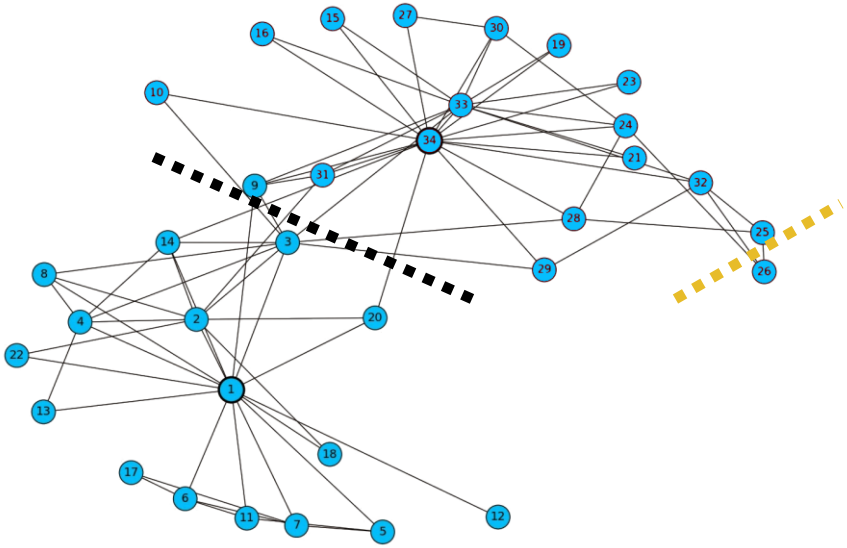
Define communities in terms of sets of nodes which are reachable from each other



- If *a* and *b* belong to a **strongly connected component** then there must be a path from a → b and a path from b → a
- A **weakly connected component** is a set of nodes that **would be** strongly connected, if the graph were undirected

2. Graph cuts

What is the **Ratio Cut** cost of the following two cuts?



$$\text{Ratio Cut}(\cdots) = \frac{1}{2}\left(\frac{3}{33} + \frac{3}{1}\right) = 1.54545$$

$$\text{Ratio Cut}(\cdots) = \frac{1}{2}\left(\frac{9}{16} + \frac{9}{18}\right) = 0.53125$$

3. Clique percolation

- Clique percolation searches for “cliques” in the network of a certain size (K). Initially each of these cliques is considered to be its own community
- If two communities share a (K-1) clique in common, they are merged into a single community
- This process repeats until no more communities can be merged

```
1. Given a clique size K
2. Initialize every K-clique as its own community
3. While (two communities I and J have a (K-1)-clique in common):
4.     Merge I and J into a single community
```

Week 3

- Clustering & Community detection – understand the basics of the different algorithms
  - Given some features, know when to apply PCA vs. K-means vs. hierarchical clustering
  - Given some networks, know when to apply clique percolation vs. graph cuts vs. connected components

CSE 258

Web Mining and Recommender Systems

Week 4

Definitions

Or equivalently...

$$R = \overbrace{\begin{pmatrix} 1 & 0 & \cdots & 1 \\ 0 & 0 & & 1 \\ \vdots & & \ddots & \vdots \\ 1 & 0 & \cdots & 1 \end{pmatrix}}^{\text{items}} \Bigg\} \text{users}$$

R_u = binary representation of items purchased by *u*

$R_{.,i}$ = binary representation of users who purchased *i*

$$I_u =$$

$$U_i =$$

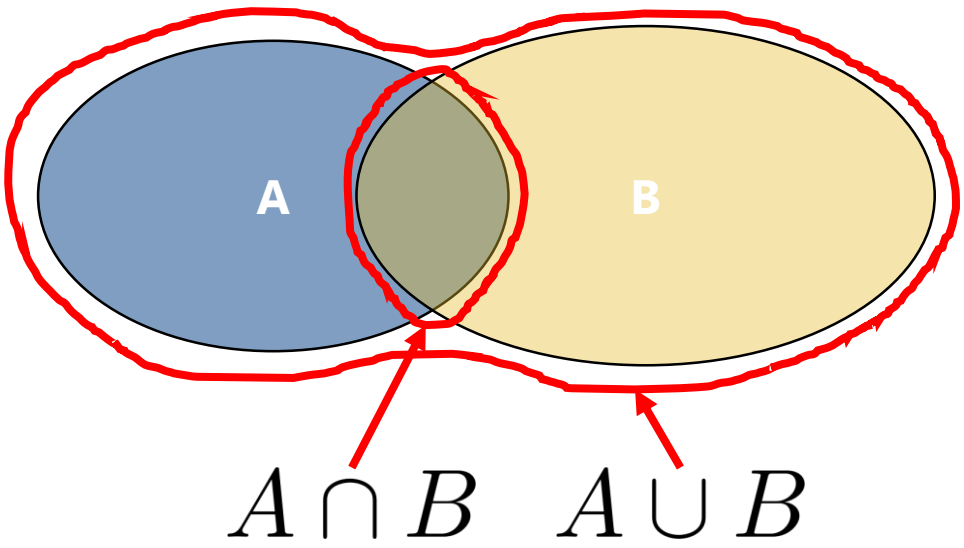
Recommender Systems Concepts

- How to represent rating / purchase data as sets/matrices
- Similarity measures (Jaccard, cosine, Pearson correlation)
- Very basic ideas behind latent factor models

Jaccard similarity

$$\text{Jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

$$\text{Jaccard}(U_i, U_j) = \frac{|U_i \cap U_j|}{|U_i \cup U_j|}$$



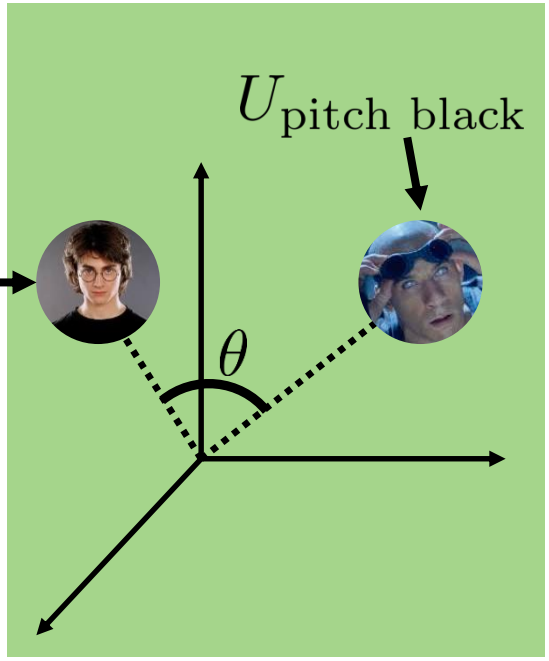
→ Maximum of 1 if the two users purchased **exactly the same** set of items
(or if two items were purchased by the same set of users)

→ Minimum of 0 if the two users purchased **completely disjoint** sets of items
(or if the two items were purchased by completely disjoint sets of users)

Cosine similarity

$$\text{Cosine}(A, B) = \frac{A \cdot B}{\|A\| \|B\|}$$

$$\theta = \cos^{-1}\left(\frac{A \cdot B}{\|A\| \|B\|}\right)$$



(vector representation of users who purchased harry potter)

$$\cos(\theta) = 1$$

(theta = 0) → *A* and *B* point in exactly the same direction

$$\cos(\theta) = -1$$

(theta = 180) → *A* and *B* point in opposite directions (won't actually happen for 0/1 vectors)

$$\cos(\theta) = 0$$

(theta = 90) → *A* and *B* are orthogonal

Pearson correlation

Compare to the cosine similarity:

Pearson similarity (between users):

items rated by both users

average rating by user *v*

$$\mathrm{Sim}(u,v) = \frac{\sum_{i \in I_u \cap I_v} (R_{u,i} - \bar{R}_u)(R_{v,i} - \bar{R}_v)}{\sqrt{\sum_{i \in I_u \cap I_v} (R_{u,i} - \bar{R}_u)^2 \sum_{i \in I_u \cap I_v} (R_{v,i} - \bar{R}_v)^2}}$$

Cosine similarity (between users):

$$\mathrm{Sim}(u,v) = \frac{\sum_{i \in I_u \cap I_v} R_{u,i} R_{v,i}}{\sqrt{\sum_{i \in I_u \cap I_v} R_{u,i}^2 \sum_{i \in I_u \cap I_v} R_{v,i}^2}}$$

Rating prediction

$$f(u, i) = \alpha + \beta_u + \beta_i$$

user item

how much does this user tend to rate things above the mean?

does this item tend to receive higher ratings than others

e.g.

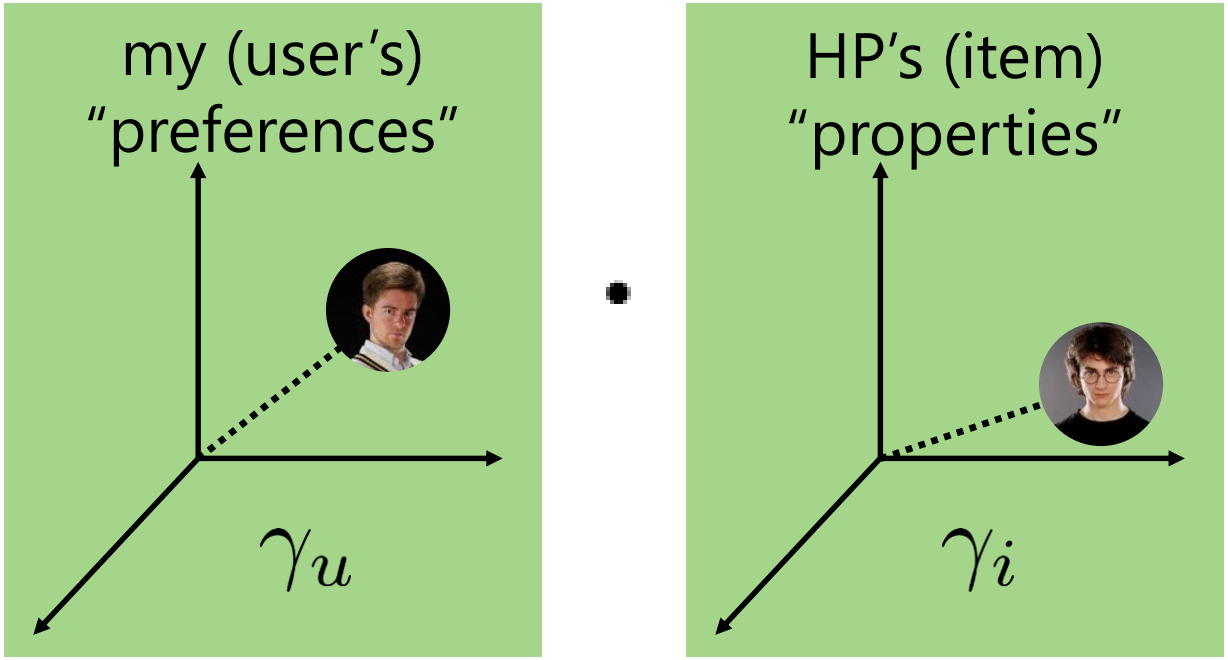
$\alpha = 4.2$

$\beta_{\text{pitch black}} = -0.1$

$\beta_{\text{julian}} = -0.2$

Latent-factor models

$$f(u,i) = \alpha + \beta_u + \beta_i + \gamma_u \cdot \gamma_i$$



CSE 258

Web Mining and Recommender Systems

Last year's midterm

Section 1: Regression and Ranking (7 marks)

Unless specified otherwise the questions in this section are each worth **1 mark**.

The following is a list of Vin Diesel's recent films:

No.	Title	Year	IMDB score	MPAA rating	length in minutes	classifier score
1	XXX: The Return of Xander Cage	2017	5.6	PG-13	110	-0.5
2	Billy Lynn's Long Halftime Walk	2016	6.6	R	113	-2.1
3	The Last Witch Hunter	2015	6.3	PG-13	106	-3.2
4	Furious 7	2015	7.4	PG-13	137	4.8
5	Guardians of the Galaxy	2014	8.1	PG-13	121	2.2
6	Riddick	2013	6.4	R	119	-1.2
7	Fast & Furious 6	2013	7.2	PG-13	130	-0.8
8	Fast Five	2011	7.3	PG-13	131	1.2
9	Fast & Furious	2009	6.6	PG-13	107	0.1
10	The Fast and the Furious: Tokyo Drift	2006	6.0	PG-13	104	-0.3

1. Suppose you train a regressor of the following form to predict the IMDB score:

$$\text{IMDB score} \simeq \theta_0 + \theta_1[\text{'Fast' in title}] + \theta_2[\text{'R' rated}] + \theta_3[\text{length in minutes}]$$

What would be the feature representation of the first two movies?

1:
2:

2. After training the above regressor you obtain $\theta = (1.5, 0.05, -0.25, 0.05)$. What would you predict would be the IMDB score of Vin Diesel's *next* film, *The Fate of The Furious* (released 2017, PG-13, 140 minutes long). You can write down an expression rather than the exact value:

A:

Section 1: Regression and Ranking (7 marks)

Unless specified otherwise the questions in this section are each worth **1 mark**.

The following is a list of Vin Diesel's recent films:

No.	Title	Year	IMDB score	MPAA rating	length in minutes	classifier score
1	XXX: The Return of Xander Cage	2017	5.6	PG-13	110	-0.5
2	Billy Lynn's Long Halftime Walk	2016	6.6	R	113	-2.1
3	The Last Witch Hunter	2015	6.3	PG-13	106	-3.2
4	Furious 7	2015	7.4	PG-13	137	4.8
5	Guardians of the Galaxy	2014	8.1	PG-13	121	2.2
6	Riddick	2013	6.4	R	119	-1.2
7	Fast & Furious 6	2013	7.2	PG-13	130	-0.8
8	Fast Five	2011	7.3	PG-13	131	1.2
9	Fast & Furious	2009	6.6	PG-13	107	0.1
10	The Fast and the Furious: Tokyo Drift	2006	6.0	PG-13	104	-0.3

Next, you train a Support Vector Machine to predict the binary outcome 'IMDB score ≥ 7.0' using the same features. Suppose the classifier produces the scores $(X_i \cdot \theta)$ shown in the right column of the table.

3. What is the accuracy, and the Balanced Error Rate of this classifier?

A:

4. What is the classifier's precision and recall?

A:

5. Perhaps you would like to add the 'year' variable to your classifier. Assuming a simple model with no regularizer, show that using the feature [year] is equivalent to using the feature [year – 2006].

A:

6. (Hard) Briefly explain why these two representations would *not* be equivalent when training a model with a regularizer (e.g. $\|\theta\|_2^2$) (**2 marks**).

A:

Section 2: Classification and Diagnostics (6 marks)

Each of the following questions is worth **2 marks**.

Suppose you are trying to train a classifier to detect whether bicycle frames will catastrophically fail during five years of use (i.e., $y_i = 1$ if the i^{th} bicycle fails). You build a dataset containing features about 10,000 bicycle frames manufactured between 1978 and 2012 (weight, material, manufacture year, etc.) and whether or not they failed. You partition the dataset into a training and validation set using a 50%/50% split.

After trying several different classifiers on your data and measuring their error (percentage of incorrect classifications), you obtain some unexpected results. Briefly explain a possible reason for the results and suggest a possible solution.

7. You obtain low (1%) error on your training set but even *lower* error (0.5%) on your validation set.

Diagnosis:
Solution:

Section 2: Classification and Diagnostics (6 marks)

Each of the following questions is worth **2 marks**.

Suppose you are trying to train a classifier to detect whether bicycle frames will catastrophically fail during five years of use (i.e., $y_i = 1$ if the i^{th} bicycle fails). You build a dataset containing features about 10,000 bicycle frames manufactured between 1978 and 2012 (weight, material, manufacture year, etc.) and whether or not they failed. You partition the dataset into a training and validation set using a 50%/50% split.

After trying several different classifiers on your data and measuring their error (percentage of incorrect classifications), you obtain some unexpected results. Briefly explain a possible reason for the results and suggest a possible solution.

8. You build an accurate classifier with only 1% error on your training set, and 1.5% error on your validation set. However, when you deploy the system, it fails to identify any instances of catastrophic failure.

Diagnosis:
Solution:

Section 2: Classification and Diagnostics (6 marks)

Each of the following questions is worth **2 marks**.

Suppose you are trying to train a classifier to detect whether bicycle frames will catastrophically fail during five years of use (i.e., $y_i = 1$ if the i^{th} bicycle fails). You build a dataset containing features about 10,000 bicycle frames manufactured between 1978 and 2012 (weight, material, manufacture year, etc.) and whether or not they failed. You partition the dataset into a training and validation set using a 50%/50% split.

After trying several different classifiers on your data and measuring their error (percentage of incorrect classifications), you obtain some unexpected results. Briefly explain a possible reason for the results and suggest a possible solution.

9. Suppose that to fix the above issue, you want to adjust a logistic regression-based classifier so that it gives 100 times as much weight to false negatives (bicycles that fail but were predicted not to) as to false positives. How would you adjust the objective function to achieve this? Recall that the original objective for logistic regression is

$$\sum_{y_i=1} \log \sigma(X_i \cdot \theta) + \sum_{y_i=0} \log(1 - \sigma(X_i \cdot \theta))$$

A:

Section 3: Clustering / Communities (8 marks)

The following questions are concerned with the *K-means* algorithm (see pseudocode at the end of the exam). Each question is worth **2 marks**.

Algorithm 1 K-means

Initialize every cluster to contain a random set of points
while cluster assignments change between iterations **do**
  Assign each X_i to its nearest centroid
  Update each centroid to be the mean of points assigned to it

---

10. Explain why the algorithm provided in the pseudocode will eventually converge (i.e., terminate).

A:

Section 3: Clustering / Communities (8 marks)

The following questions are concerned with the *K-means* algorithm (see pseudocode at the end of the exam). Each question is worth **2 marks**.

Algorithm 1 K-means

Initialize every cluster to contain a random set of points
while cluster assignments change between iterations **do**
  Assign each X_i to its nearest centroid
  Update each centroid to be the mean of points assigned to it

---

11. The K-Means algorithm will in general converge to a local optimum rather than a global one. Draw a simple 2-d example, containing a set of points and clusters, such that the solution is *not* optimal but for which the algorithm would not make further progress.

A:

Section 3: Clustering / Communities (8 marks)

The following questions are concerned with the *K-means* algorithm (see pseudocode at the end of the exam). Each question is worth **2 marks**.

Algorithm 1 K-means

Initialize every cluster to contain a random set of points
while cluster assignments change between iterations **do**
  Assign each X_i to its nearest centroid
  Update each centroid to be the mean of points assigned to it

---

12. Suggest simple modifications to the k-means algorithm that might increase its chances of finding a good solution.

A:

13. Compared to PCA, K-means will work well for different types of clusters. Give three examples of 2-d clustered data where (a) K-means will perform *better* than PCA (in terms of reconstruction error); (b) K-means will perform *worse* than PCA; and (c) K-means and PCA will both perform poorly.

A:

Section 4: Recommender Systems (5 marks)

Unless specified otherwise the questions in this section are each worth **1 mark**.

On a popular movie streaming website, a few users have watched the following recent movies:

Movie	Watched?				Rated?			
	Caroline	Mengting	Ruining	Zachary	Caroline	Mengting	Ruining	Zachary
XXX: Return of Xander Cage	1	1	0	1	5	1		2
La La Land	1	1	1	1	5	2	2	2
John Wick 2	1	1	1	0	4	2	1	
Rogue One	0	0	1	1			5	1
Resident Evil	1	0	0	1	4			1

14. Using 'watched' data: Which two users are *most similar* in terms of their Jaccard similarity (write down all pairs in case of a tie)?

A:

15. Which two *items* are most similar in terms of their Jaccard similarity?

A:

Section 4: Recommender Systems (5 marks)

Unless specified otherwise the questions in this section are each worth **1 mark**.

On a popular movie streaming website, a few users have watched the following recent movies:

Movie	Watched?				Rated?			
	Caroline	Mengting	Ruining	Zachary	Caroline	Mengting	Ruining	Zachary
XXX: Return of Xander Cage	1	1	0	1	5	1		2
La La Land	1	1	1	1	5	2	2	2
John Wick 2	1	1	1	0	4	2	1	
Rogue One	0	0	1	1			5	1
Resident Evil	1	0	0	1	4			1

16. Which two users are most similar in terms of their *ratings*, based on their Pearson correlation (defined below)?

A:

$$\text{Sim}(u,v) = \frac{\sum_{i \in I_u \cap I_v} (R_{u,i} - \bar{R}_u)(R_{v,i} - \bar{R}_v)}{\sqrt{\sum_{i \in I_u \cap I_v} (R_{u,i} - \bar{R}_u)^2 \sum_{i \in I_u \cap I_v} (R_{v,i} - \bar{R}_v)^2}}$$

(R = rating matrix, I_u = items rated by user u, $\bar{R}_u$ = average rating by user u)

17. (Hard) Show that a latent factor model of the form

$$\text{rating}(u, i) = \alpha + \beta_u + \beta_i$$

is linear in its parameters ($\theta = (\alpha; \beta_u; \beta_i)$) but that a model of the form

$$\text{rating}(u, i) = \alpha + \beta_u + \beta_i + \gamma_u \cdot \gamma_i$$

is not linear in its parameters ($\theta = (\alpha; \beta_u; \beta_i; \gamma_u; \gamma_i)$) (recall the definition of linearity: $r_{\theta_1+\theta_2}(u, i) = r_{\theta_1}(u, i) + r_{\theta_2}(u, i)$) (**2 marks**).

A:

CSE 258

Web Mining and Recommender Systems

Fall 2015 midterm

Last year's midterm

Section 1: Regression

Feature design

Suppose we collected the following data about businesses from *Google Local*:

ID	Name	Av. Rating	Price	Address	latitude/longitude	hours
0	T C's Referee Sports Bar	5.0	$$	Sioux Falls, SD 57106	43.529, -96.792	m-f/11am-10pm, s-s/11am-1am
1	Old Chicago	3.0	$$	Beaverton, OR 97006	45.535, -122.862	m-f/11am-1am
2	Sabatinos Italian Kitchen	4.0	$$$	Arlington, MA 02474	42.406, -71.143	m-f/10am-10pm, s-s/10am-9pm
3	Oakville Grocery	4.5	$	Healdsburg, CA 95448	25.063, 121.524	mon-sun/9am-5pm
4	Hog Wild Pit BBQ	3.5	$$	Wichita, KS 67213	37.681, -97.389	mon-sun/11am-8pm

1. Suppose we wanted to train a personalized model that predicted the rating I would give to a business based on the population-level average and price, i.e.,

$$\text{my rating} = \theta_0 + \theta_1[\text{average rating}] + \theta_2[\text{price}].$$

Write down the complete feature matrix (in the space below) that you would use to solve the above equation (1 mark):

$$y = \begin{bmatrix} & \\ & \\ & \\ & \\ & \end{bmatrix} \theta$$

Last year's midterm

2. Write down the predictions that would be obtained for the five businesses if using the features above if the parameters were $\theta = [0.1, 1.0, -0.2]^T$ (1 mark)

ID	Name	Predicted Rating	Q4 answer	Q5 answer
0	T C's Referee Sports Bar			
1	Old Chicago			
2	Sabatinos Italian Kitchen			
3	Oakville Grocery			
4	Hog Wild Pit BBQ			

3. The opening hours might also influence my preferences. How would you construct useful features for the above businesses, if I have a preference toward *businesses that are open late*? Using your representation write down (in the table above) the features corresponding to each business (1 mark):

A:

Last year's midterm

4. How would you incorporate opening-hour features if my preferences are toward *businesses that are open outside of work hours (i.e., mon-fri/9am-5pm)*? Write down the features corresponding to each business (1 mark):

A:

5. Finally, suppose I want to model how people's preferences change as a function of geography (based on 1,000 U.S. businesses including the five above), i.e.,

$$\text{average rating} = x(\text{geographical features}) \cdot \theta$$

How might you use the features available above (e.g. address or latitude/longitude) to model such geographical trends (1 mark)? (describe your solution, rather than writing down the actual features)

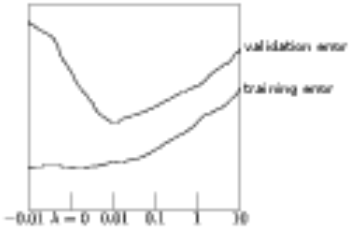
Last year's midterm

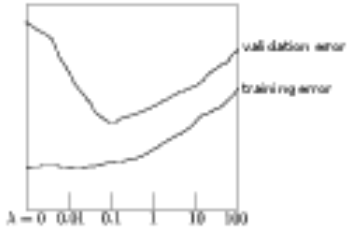
Diagnostics

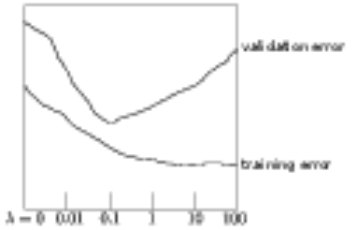
6. Suppose we trained our model above by minimizing the *regularized mean squared error*, i.e.,

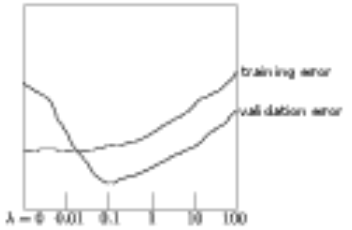
$$\underset{\theta}{\operatorname{argmin}} \|y - X\theta\|_2^2 + \lambda\|\theta\|_2^2$$

Suppose that we split our data into training, validation, and test sets (and that we do so randomly, given plenty of data). Which of the plots below could correspond to the performance (i.e., MSE) on the training and validation sets? For each that could *not*, briefly explain why below (1 mark).









Last year's midterm

(hard) Suppose you are trying to predict star ratings using some regression model (e.g. $\alpha + \beta_u + \beta_i$). You figure that since the output of your model is a real number, while the labels themselves are *integers* (i.e., 1, 2, 3, 4, or 5), that you might simply round the output to the nearest integer to improve your predictor. You perform a quick check and find that when your model outputs the number a, the correct answer is $\lfloor a \rfloor$ with probability $\lceil a \rceil - a$, and $\lceil a \rceil$ with probability $a - \lfloor a \rfloor$ (e.g. if it outputs 4.2 then the correct answer is 4 80% of the time and 5 20% of the time; recall that $\lfloor \cdot \rfloor$ and $\lceil \cdot \rceil$ are the floor and ceiling function).

7. Based on the above description, would rounding be expected to increase or decrease the MSE, or have no effect? Explain your answer (2 marks).

   A:

8. What effect would the above rounding procedure have if we were trying to optimize the mean *absolute* error (MAE) instead of the MSE (2 marks)? Explain.

   A:

Last year's midterm

Section 2: Classification

The following is a list of Vin Diesel's films:

No.	Title	Year	IMDB rating	MPAA rating	length (minutes)
1	The Last Witch Hunter	2015	6.3	PG-13	106
2	Furious 7	2015	7.4	PG-13	137
3	Guardians of the Galaxy	2014	8.1	PG-13	121
4	Riddick	2013	6.4	R	119
5	Fast & Furious 6	2013	7.2	PG-13	130
6	Fast Five	2011	7.3	PG-13	131
7	Fast & Furious	2009	6.6	PG-13	107
8	The Fast and the Furious: Tokyo Drift	2006	6.0	PG-13	104
9	The Pacifier	2005	5.5	PG	95
10	The Chronicles of Riddick	2004	6.7	PG-13	119
11	xXx	2002	5.8	PG-13	124
12	The Fast and the Furious	2001	6.7	PG-13	106
13	Pitch Black	2000	7.1	R	109
14	The Iron Giant	1999	8.0	PG	86
15	Saving Private Ryan	1998	8.6	R	169

You hear a rumor that Vin Diesel has a new film coming out that is (A) Over two hours long (B) Rated PG-13 (C) Has the word "Furious" in the title. Let's try to estimate the probability that it will (D) have an IMDB rating of 7.0 or above.

9. Based on the data above (and not making any other assumptions) write down the probability

$$p(D|A \wedge B \wedge C)$$

(1 mark) A:

Last year's midterm

9. Based on the data above (and not making any other assumptions) write down the probability

$$p(D|A \wedge B \wedge C)$$

(1 mark) A:

10. The above probability may be unreliable as it is based on very few observations that exhibit the required features. So, we'll try to decide whether D is likely to be true or not following the Naïve Bayes assumption. Write down all of the terms involved and finally the probability ratio, and the conclusion you draw as a result (2 marks).

A:

11. Can you comment on the appropriateness of the naïve bayes assumption for this task (i.e., predicting IMDB ratings based on movie features) (1 mark)?

A:

Last year's midterm

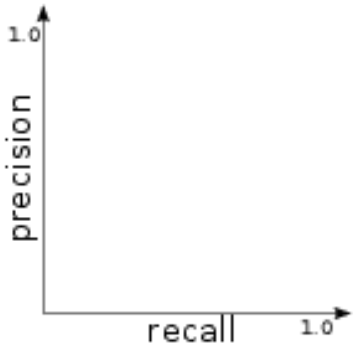
Evaluation measures

Suppose we are performing a ranking task to try and identify pages that are relevant to some particular search query, and that we achieve this by building a logistic regressor that outputs a score indicating the probability that a page is relevant. Suppose the scores we obtain are the following:

page id	score	actually relevant?
0	0.78	yes
1	0.25	no
2	0.36	yes
3	0.18	no
4	0.01	no
5	0.95	yes
6	0.92	yes
7	0.11	no
8	0.20	no
9	0.56	no

12. Complete the table below by ranking pages in decreasing order of confidence. (Roughly) plot the precision against the recall to the right of the table (3 marks).

page id	confidence	actually relevant?	precision@k	recall@k
5	0.95	yes		



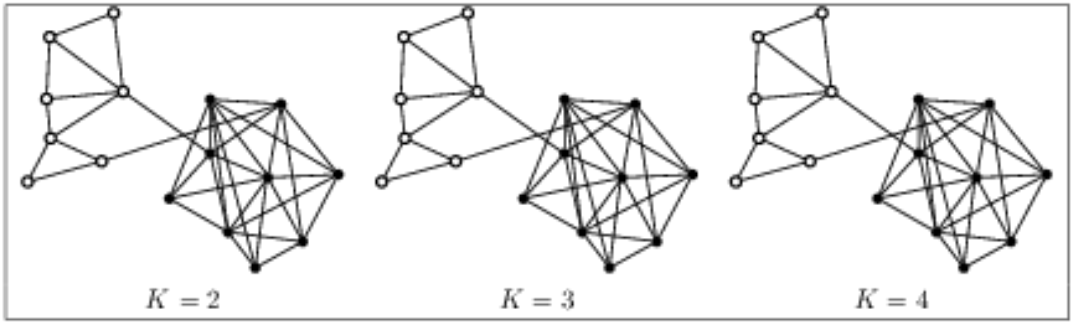
Last year's midterm

Section 3: Communities & clustering

13. Suppose a social network is divided into the two communities shown below (filled vs. unfilled nodes). If we wanted an algorithm to find these communities automatically, which of *ratio cuts* versus *normalized cuts* would be more appropriate and why (1 mark)?

A:

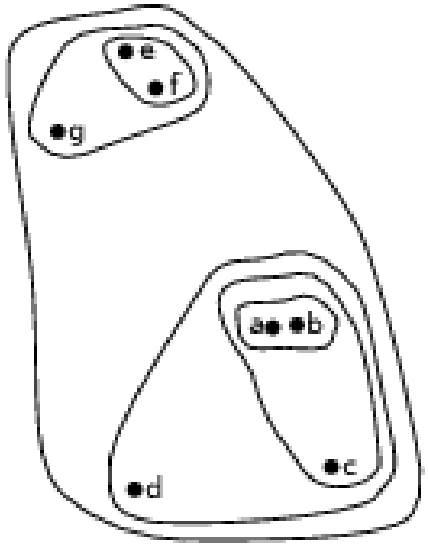
14. What would be the result of running *clique percolation* on the graphs below (3 marks)? Circle the communities that would be found directly on the graphs.

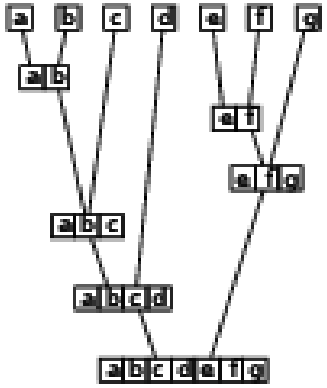


Last year's midterm

15. Suppose you ran *hierarchical clustering* on the points below, resulting in the dendrogram shown in the center. How would you use the output of this algorithm (i.e., the clusters/dendrogram) to generate useful feature representations for the original points? Write your features for the 7 points below (1 mark).

A:





$X =$

Last year's midterm

Algorithm design

16. Suppose you wanted to design a system to estimate what tip a prospective fare would give for a taxi ride in San Diego. Describe below what data and features you would collect to estimate this value, and what techniques you would use to solve the task (3 marks).

A:

CSE 258

Web Mining and Recommender Systems

Spring 2015 midterm

Spring 2015 midterm

Section 1: Regression

Q1: Restaurants & ratings (10 marks)

Suppose we collected the following data about restaurants from *Yelp!*:

Name	Average Rating	Takes reservations?	Take-out?	Price	Good for
Oceana Coastal Kitchen	4.5	Yes	No	$$$	Breakfast
Beyer Deli	5.0	No	Yes	$	Lunch
Werewolf	4.5	Yes	Yes	$$	Brunch
C Level	4.0	No	Yes	$$	Lunch, Dinner
Cucina Urban	4.5	Yes	Yes	$$	Dinner

and that from this data we want to estimate

$$\text{av. rating} \simeq \theta_0 + \theta_1[\text{takes reservations}] + \theta_2[\text{has take-out}] + \theta_3[\text{price}]$$

Spring 2015 midterm

Name	Average Rating	Takes reservations?	Take-out?	Price	Good for
Oceana Coastal Kitchen	4.5	Yes	No	$$$	Breakfast
Beyer Deli	5.0	No	Yes	$	Lunch
Werewolf	4.5	Yes	Yes	$$	Brunch
C Level	4.0	No	Yes	$$	Lunch, Dinner
Cucina Urban	4.5	Yes	Yes	$$	Dinner

and that from this data we want to estimate

$$\text{av. rating} \simeq \theta_0 + \theta_1[\text{takes reservations}] + \theta_2[\text{has take-out}] + \theta_3[\text{price}]$$

1. What is the average rating across all restaurants (1 mark)? A:

2. What is the Mean Squared Error of the a predictor that just predicts the average rating for all items (1 mark)? A:

3. Suppose we'd like to write down the above expression for the rating in the form $y \simeq X\theta$. Complete the following equation to do so:

$$\begin{bmatrix} 4.5 \\ \\ \\ \\ \\ \end{bmatrix} \simeq \begin{bmatrix} 1 & 1 & 0 & 3 \\ & & & \\ & & & \\ & & & \\ & & & \end{bmatrix} \theta$$

(1 mark)

4. In the expression $y \simeq X\theta$, which term encodes the labels, which term encodes the features, and which term encodes the parameters (1 mark)? labels: features: parameters:

Spring 2015 midterm

Name	Average Rating	Takes reservations?	Take-out?	Price	Good for
Oceana Coastal Kitchen	4.5	Yes	No	$$$	Breakfast
Beyer Deli	5.0	No	Yes	$	Lunch
Werewolf	4.5	Yes	Yes	$$	Brunch
C Level	4.0	No	Yes	$$	Lunch, Dinner
Cucina Urban	4.5	Yes	Yes	$$	Dinner

and that from this data we want to estimate

$$\text{av. rating} \simeq \theta_0 + \theta_1[\text{takes reservations}] + \theta_2[\text{has take-out}] + \theta_3[\text{price}]$$

5. Suppose that after fitting our model for the rating we obtain $\theta = [7, 0.5, -1, -1]^T$. What is the interpretation of $\theta_0 = 7$ in this expression (1 mark)?

   A:

6. What is the interpretation of $\theta_3 = -1$ (1 mark)?

   A:

Spring 2015 midterm

Name	Average Rating	Takes reservations?	Take-out?	Price	Good for
Oceana Coastal Kitchen	4.5	Yes	No	$$$	Breakfast
Beyer Deli	5.0	No	Yes	$	Lunch
Werewolf	4.5	Yes	Yes	$$	Brunch
C Level	4.0	No	Yes	$$	Lunch, Dinner
Cucina Urban	4.5	Yes	Yes	$$	Dinner

and that from this data we want to estimate

$$\text{av. rating} \simeq \theta_0 + \theta_1[\text{takes reservations}] + \theta_2[\text{has take-out}] + \theta_3[\text{price}]$$

9. Suppose you wanted to incorporate the 'Good for' field (the last column of the above table) into your model. How would you represent the features in order to do so? Answer this by writing down the model you would use:

$$\text{av. rating} \simeq \theta_0 + \theta_1[\text{takes reservations}] + \theta_2[\text{has take-out}] +$$
$$\theta_3[\text{price}] + \boxed{\text{A:}}$$

2

and by completing the feature matrix using your representation:

$$X = \begin{bmatrix} 1 & 1 & 0 & 3 \\ & & & \\ & & & \end{bmatrix}$$

(2 marks)

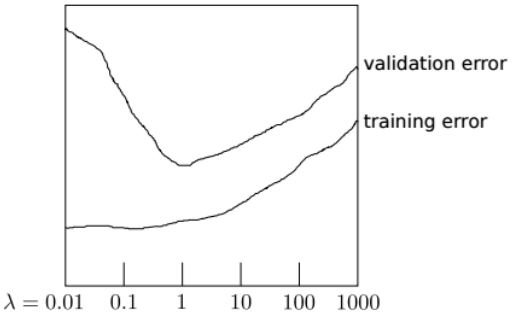
Spring 2015 midterm

Q2: Training, testing, & model selection (6 marks)

Suppose we are training regressors to minimize the regularized Mean Squared Error

$$\sum_{(x,y)\in train} \frac{1}{|train|}(y - x \cdot \theta)^2 + \lambda\|\theta\|_2^2.$$

10. Suppose that we fit some model for $\lambda \in \{0.01, 0.1, 1, 10, 100, 1000\}$ and obtain the following performance on the training and validation sets:



Which value of λ would you select based on the results above (1 mark)? $\lambda =$

11. Answer the following questions about training, validation, and test sets:

(a) What is the role of a validation set (1 mark)?

A:

(b) How does the training error typically vary with λ (1 mark)?

A:

(c) What is meant by under/over fitting? Which values of λ in the above figure correspond to maximum over/under fitting (1 mark)?

A:

Spring 2015 midterm

12. Further suppose that we consider two different feature representations (model X and model Y), and two different regularization parameters ($\lambda = 1$ and $\lambda = 10$) and obtain the following results on the training and validation sets:

model	training error	validation error
model X, $\lambda = 1$	23.34	?
model X, $\lambda = 10$	?	?
model Y, $\lambda = 1$	?	18.32
model Y, $\lambda = 10$	25.98	?

('?' indicates an unknown value).

Assuming that our training/validation/test sets are large, independent samples, is the above information enough to determine which model and which value of λ we would expect to yield the best performance

on the test set? If so, which model and which value of λ would you expect to perform best and why? Explain your answer (2 marks).

A:

Spring 2015 midterm

Q3: Fantasy novels (6 marks)

Suppose we have a database consisting of the following books:

Title	Genre	Prediction
The Circle of Sorcerers	Fantasy	True
Knights: The Eye of Divinity	Fantasy	
Superman/Batman: Sorcerer Kings	Graphic Novel	
In the Blood	Mystery	
Remains of the Day	Literature & Fiction	
Blood Song	Fantasy	
Flame Moon	Fantasy	
The Book of The Sword: A History of Daggers	History	
A Storm of Swords	Fantasy	
The Storm Book	Children's	

Further, suppose we are given the following classifier to classify Fantasy vs. non-Fantasy books:

```
if (Title contains 'Sorcerer' or 'Blood' or 'Knights' or 'Moon' or 'Storm'):
  return True
else:
  return False
```

13. What are the predictions made by this classifier? Write your answers in the last column of the table above (1 mark).

14. Of these predictions, what is the number of true positives, true negatives, false positives, and false negatives (1 mark)?

	true positive	true negative	false positive	false negative
A:				

15. What are the true positive rate (hint: TP / (TP + FN)), true negative rate, and balanced error rate (1 mark)?

	true positive rate	true negative rate	balanced error rate
A:			

16. In class we saw three approaches to classification: naïve Bayes, logistic regression, and support-vector machines. Describe one benefit of each approach compared to the other two (3 marks).

naïve Bayes:

logistic regression:

SVM:

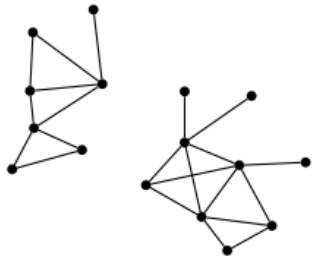
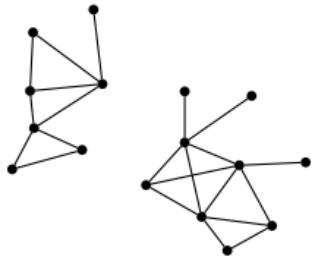
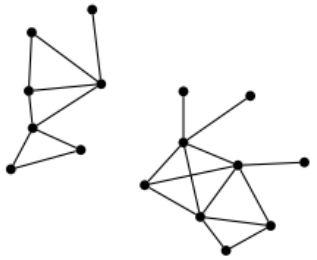
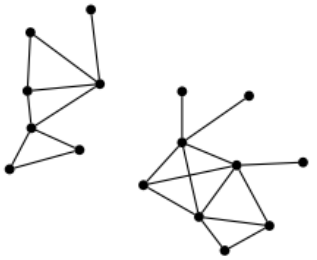
Spring 2015 midterm

Section 3: Communities & clustering

Q4: Algorithms for community detection, dimensionality reduction, and clustering

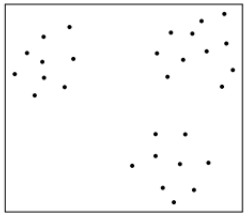
Recall three algorithms we saw in class to detect communities: connected components, ratio cut, and clique percolation (pseudocode is given as Algorithms 1, 2, and 3 at the end of the test).

17. Identify the communities that would be produced on the graphs below using these three algorithms. Circle the communities directly in the space below (some more graphs are provided on the final page in case you need to re-write your answer):

Connected components	Ratio cut (2 communities)	Clique percolation ($k = 3$)	Clique percolation ($k = 2$)
			
(1 mark)	(1 mark)	(1 mark)	(1 mark)

Spring 2015 midterm

18. Suppose we are given the following 2-dimensional data X, and wish to cluster it so as to minimize the reconstruction error ($\sum_{x \in X} \|\bar{x} - x\|_2^2$). Separate the points into three clusters such that the reconstruction error (when replacing each point by its cluster centroid) would be minimized. Draw the clusters directly in the space below (1 mark):

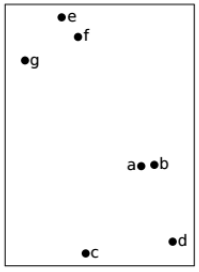


19. By replacing each point with one of the three centroids above, we have effectively 'compressed' the data, since each (2-d) point is replaced by a (1-d) integer. Another way to compress the data would be to perform Principal Component Analysis, and discard the lowest variance dimension, which would also result in a 1-d representation of the data. Out of these two possible compressed representations, which one would result in the lower reconstruction error on the above data, and why (1 mark)?

A:

20. In class we saw *hierarchical clustering*, an algorithm that works by successively joining clusters whose centroids are nearest. Psuedocode is given in Algorithm 4 over the page.

    Suppose you are given the following set of points:



Step	Clusters merged	List of clusters
0	(initialization)	$\{a\}, \{b\}, \{c\}, \{d\}, \{e\}, \{f\}, \{g\}$
1	$\{a\}$ merges with $\{b\}$	$\{a, b\}, \{c\}, \{d\}, \{e\}, \{f\}, \{g\}$
2		
3		
4		
5		
6		$\{a, b, c, d, e, f, g\}$

    If we were to perform hierarchical clustering on this data, in what order would the clusters be joined? Answer this question by completing the table above (2 marks).

CSE 258 – Lecture 10

Web Mining and Recommender Systems

Misc. questions

Misc. Qs