CSE 190 – Lecture 10

Data Mining and Predictive Analytics

Homework and midterm recap

Assignment 1

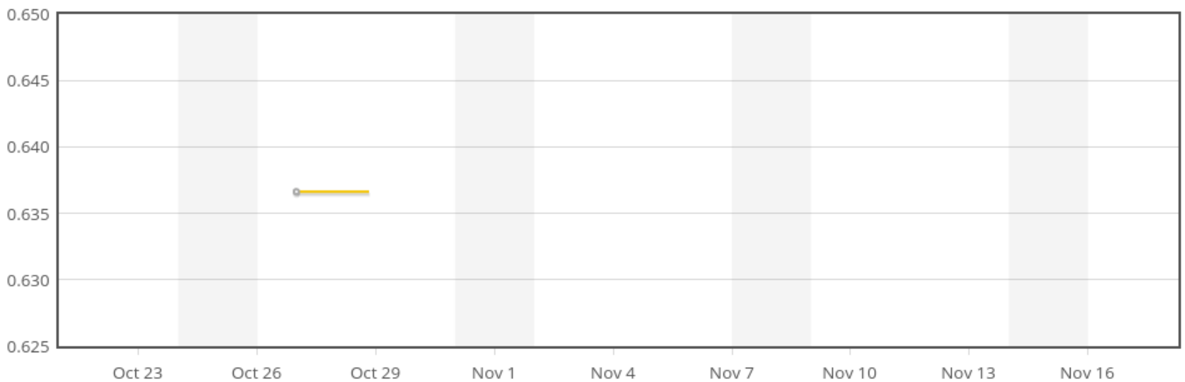
Knowledge • 4 teams

CSE 190/255 (fa15) -- Assignment 1 -- Task 1 -- Helpfulness Prediction

Tue 20 Oct 2015 — Wed 18 Nov 2015 (20 days to go)

Dashboard

Public Leaderboard - CSE 190/255 (fa15) -- Assignment 1 -- Task 1 -- Helpfulness Prediction



This leaderboard is calculated on approximately 50% of the test data.
The final results will be based on the other 50%, so the final standings may be different.

See someone using multiple accounts?
Let us know.

#	Δ1d	Team Name	Score	Entries	Last Submission UTC (Best – Last Submission)
1	—	Baseline	0.63663	1	Mon, 26 Oct 2015 23:40:54
2	new	kaizhou	0.63663	1	Tue, 27 Oct 2015 21:45:17
3	new	ShubhamSaini	0.63663	1	Wed, 28 Oct 2015 09:16:58
4	new	playground	0.63663	1	Wed, 28 Oct 2015 10:25:21

Download raw data

Midterm on Monday!

- 5:10 pm – 6:10 pm
- Closed book – but I’ll provide a similar level of basic info as in the last page of last quarter’s midterm
- I’ll run extra office hours from 9:30-11:30 tomorrow, though the TAs will have regular office hours on Friday

CSE 190 – Lecture 10

Data Mining and Predictive Analytics

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)

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

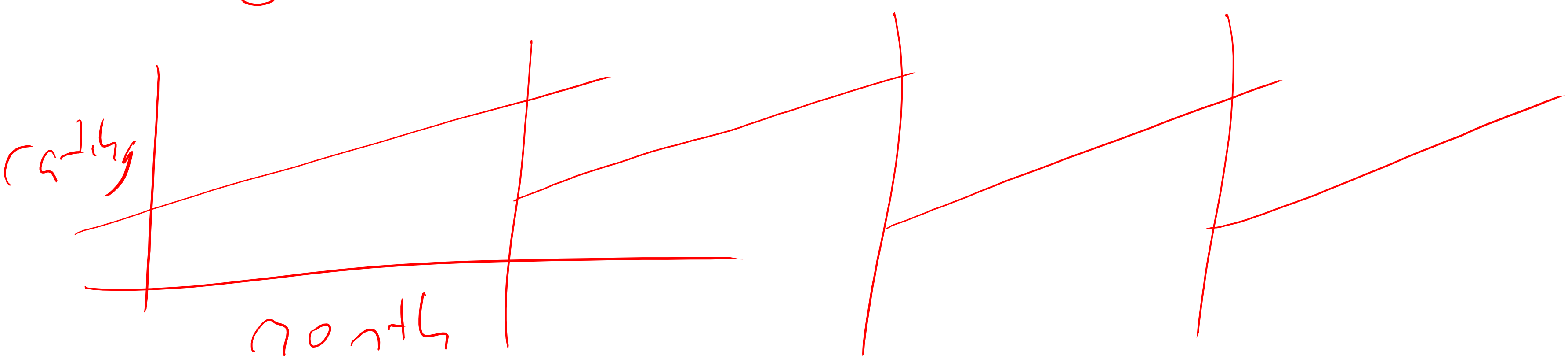
$$= \frac{1}{N}\sum_{i=1}^{N}(y_i - X_i \cdot \theta)^2$$

Representing the month as a feature

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

$$rating = \theta_0 + \theta_1 f(month)$$

jan = 1 ... Dec = 12

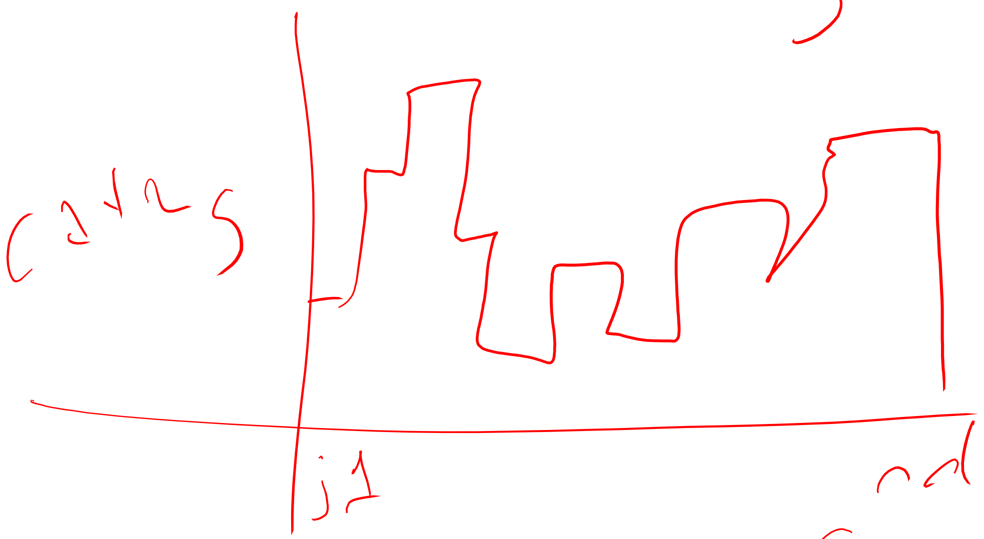


Representing the month as a feature

jan = [1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]

nov = [... 1]

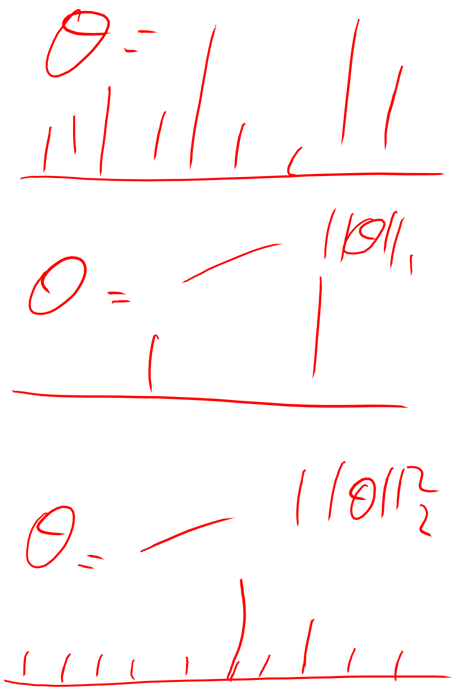
$y = \theta x$



$(X^T X)^{-1}$

Occam's razor

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



(image from personalspirituality.net)

Regularization

Regularization is the process of penalizing model complexity during training

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

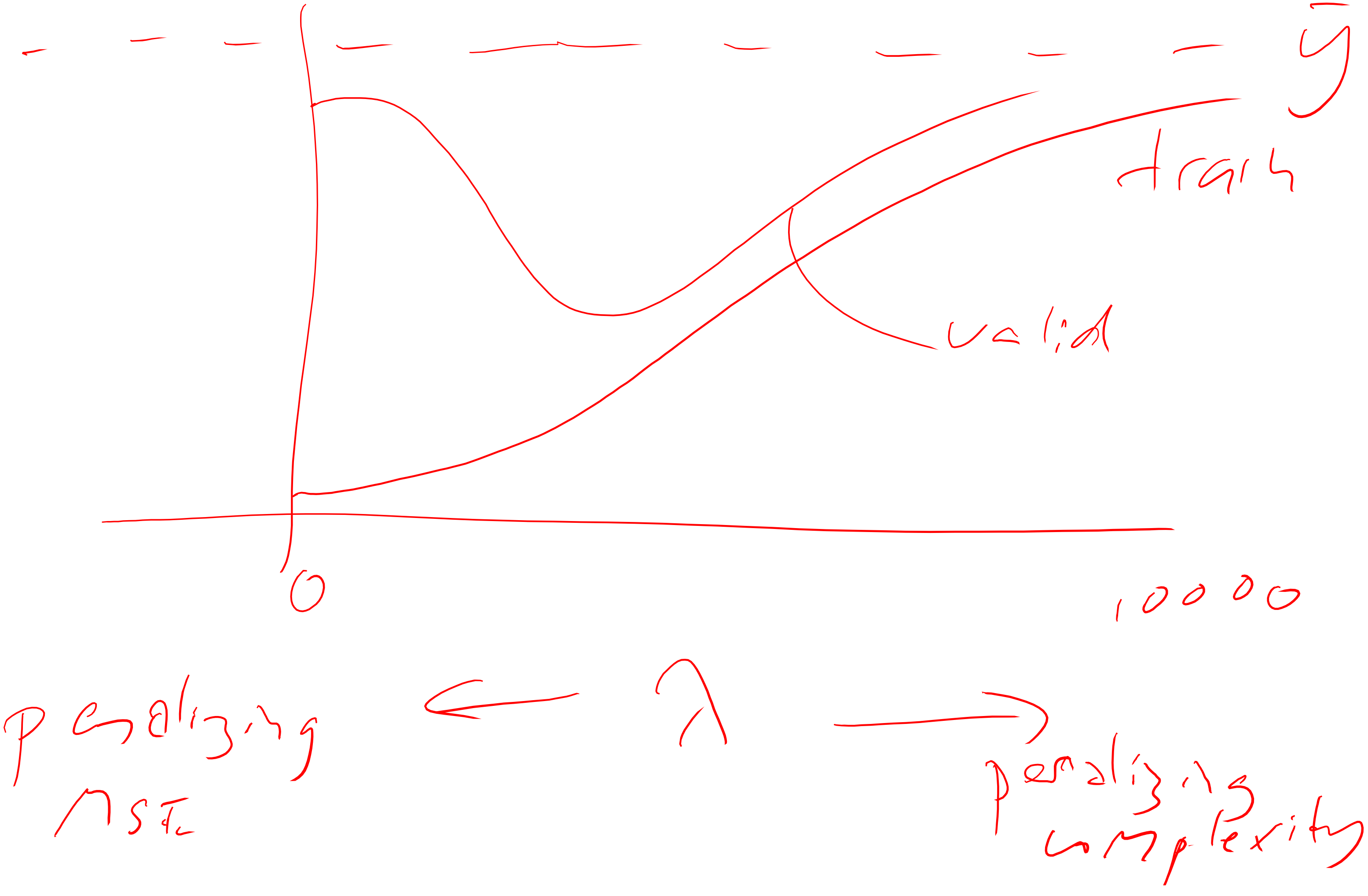
How much should we trade-off accuracy versus complexity?

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

λ increases, train increases

- 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)

  valid error $\geq$ train error

- The validation/test error will usually have a "sweet spot" between under- and over-fitting

CSE 190 – Lecture 10

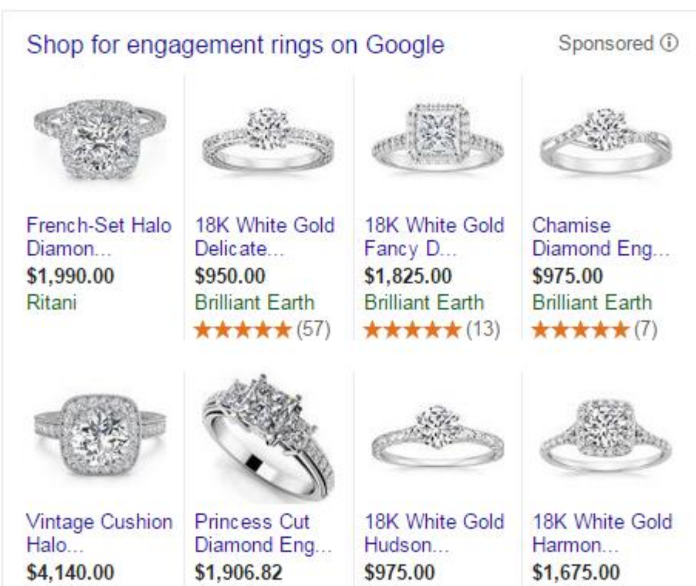
Data Mining and Predictive Analytics

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

$$X\theta = y$$

matrix of features
(data)

unknowns
(which features are relevant)

vector of outputs
(labels)

Regression vs. classification

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

$$f(\text{data}) \overset{?}{\rightarrow} \text{labels}$$

{0,1}, {True, False}

{1, … , N}

(linear) classification

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

$y_i = X_i \cdot \theta$

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

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)

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

$$\downarrow$$

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

$$=$$

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

Naïve Bayes (2 slide summary)

$$p(y \mid X) = \frac{p(y)\,p(X \mid y)}{p(X)}$$

(Bayes)

$$\frac{p(y)\prod_i p(X_i \mid y)}{p(\neg y)\prod_i p(X_i \mid \neg y)} \gtrless 1?$$

Double-counting: naïve Bayes vs Logistic Regression

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

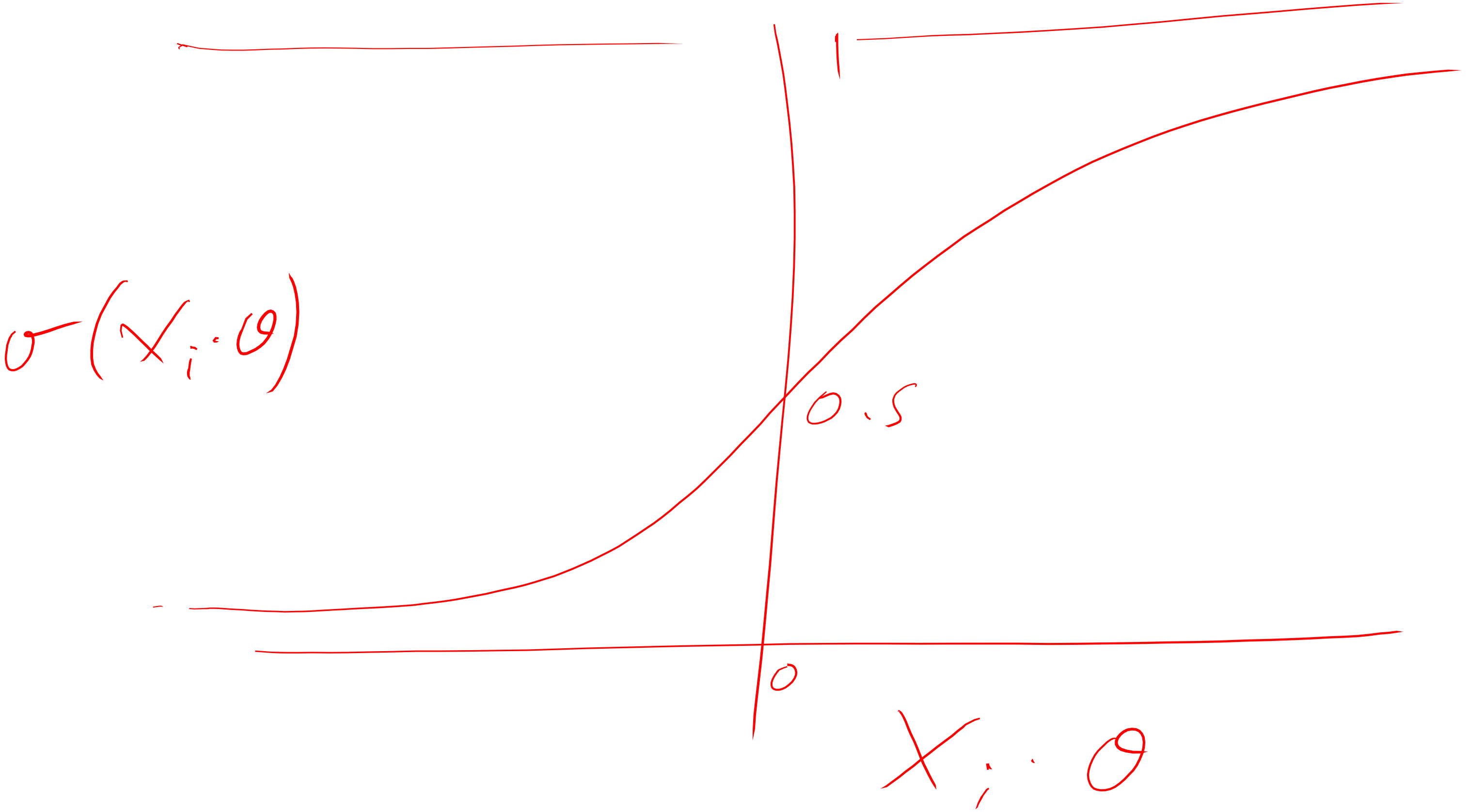
$$\text{no. of pages} = \alpha + \beta_1 \cdot \delta(\text{mentions wizards})$$

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

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

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

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

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

Logistic regression

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

obj =

(1) log (obj)

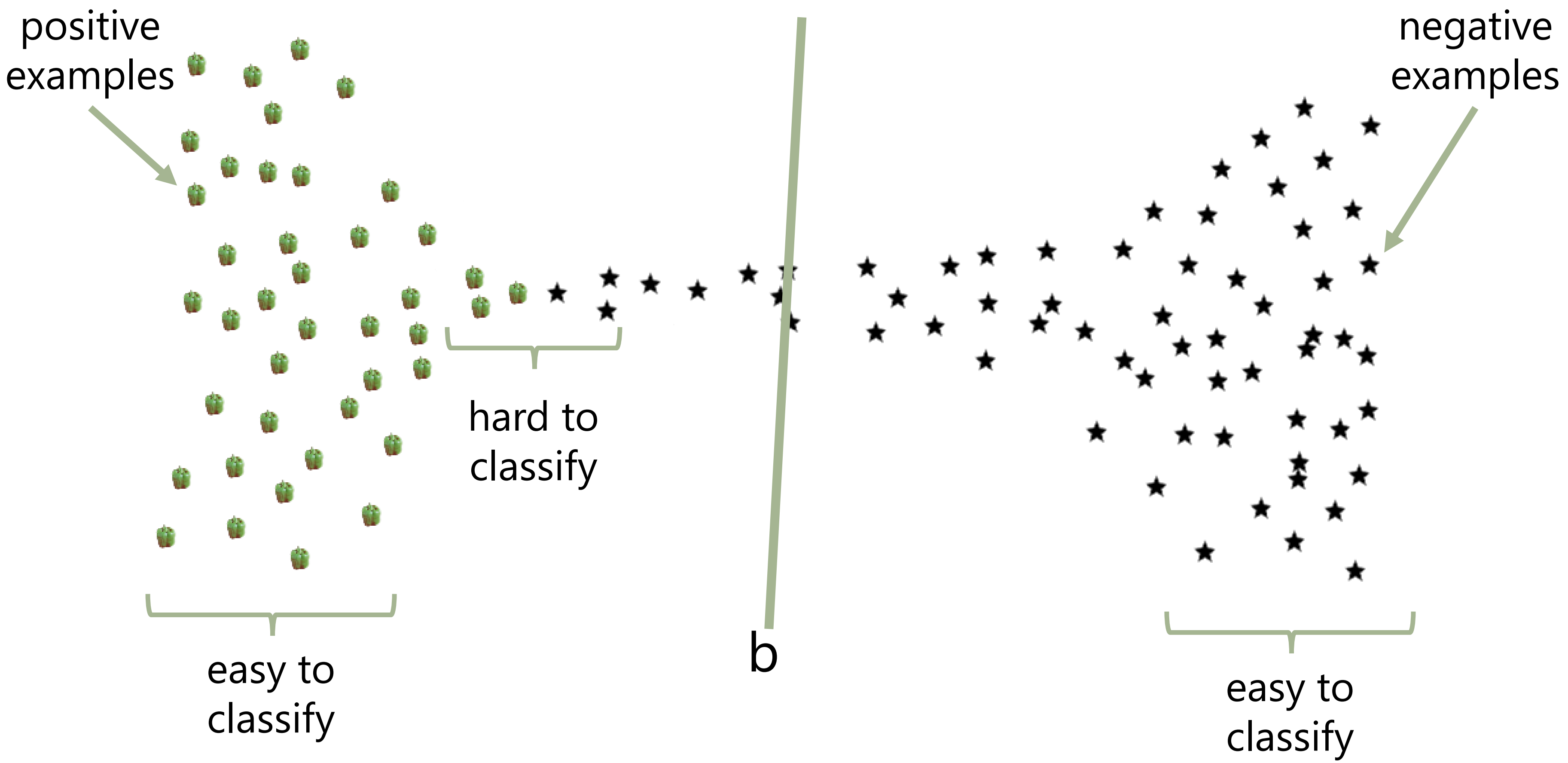
(2) subtract regularizer

(3) differentiate

(4) hillclimb

Logistic regression

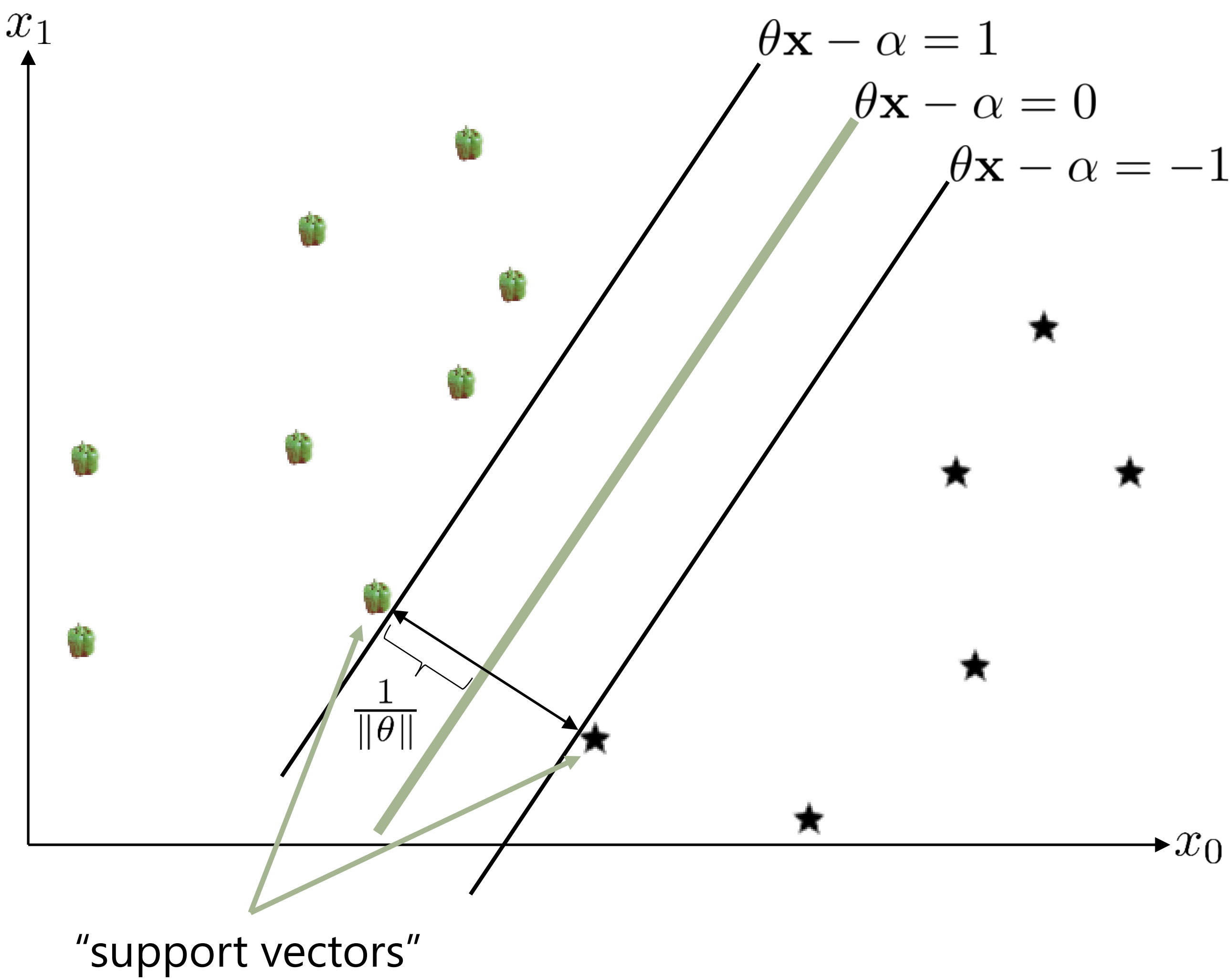
Q: Where would a logistic regressor place the decision boundary for these features?



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



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

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

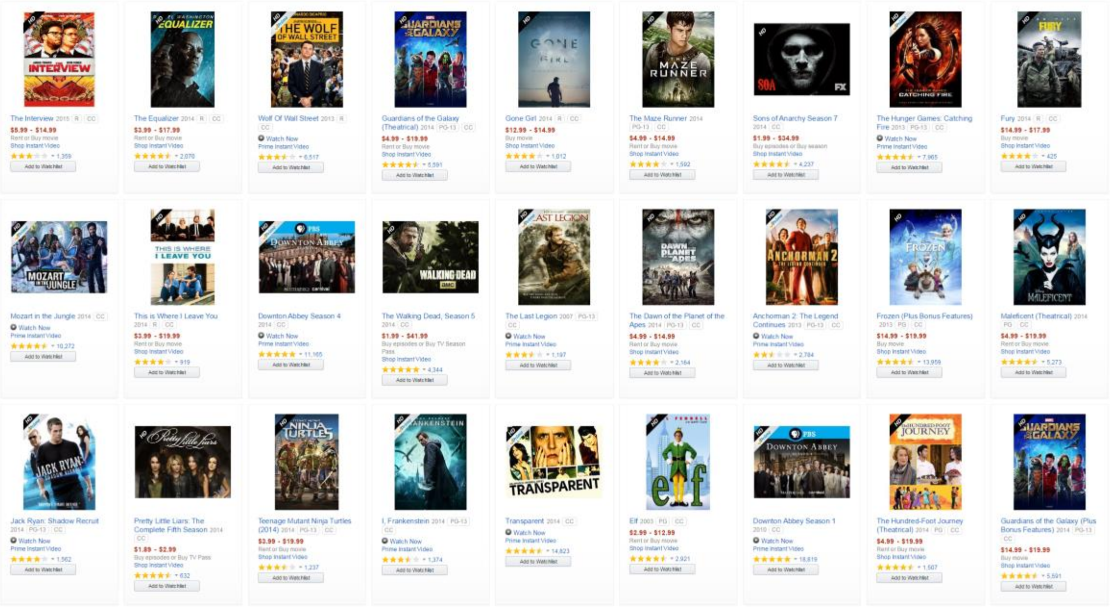
- **Naïve Bayes**
  - Probabilistic model (fits $p(label|data)$)
  - Makes a conditional independence assumption of the form $(feature_i \perp\!\!\!\perp feature_j | label)$ allowing us to define the model by computing $p(feature_i | label)$ for each feature
  - Simple to compute just by counting
- **Logistic 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)

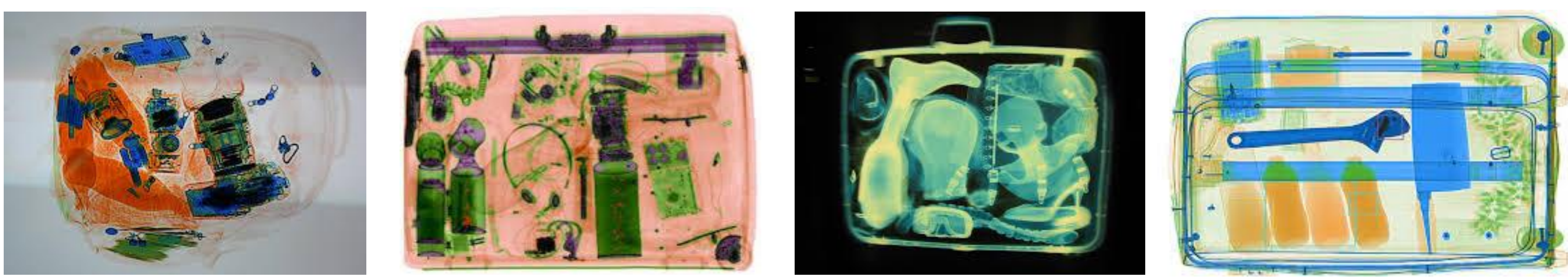
e.g. will I purchase a product? If I purchase 0.00001% of products, then a classifier which just predicts "no" everywhere is 99.99999% accurate, but not very useful



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)

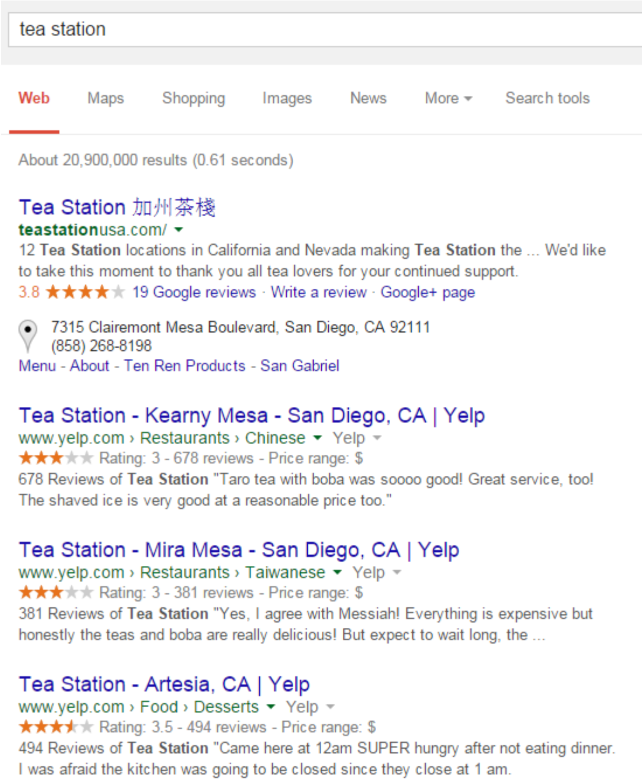


e.g. which of these bags contains a weapon?

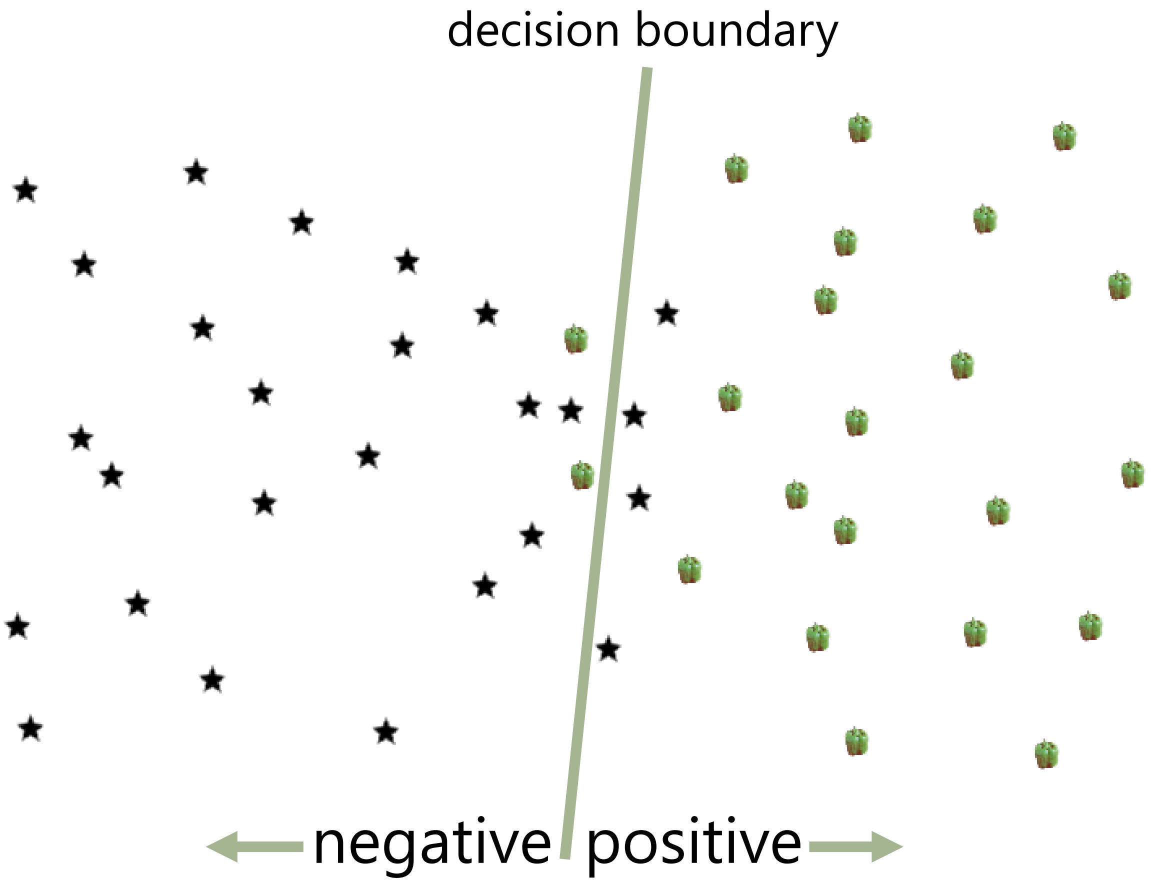
Which classifier is best?

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

		Label	
		true	false
Prediction	true	true positive	false positive
	false	false negative	true negative

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 190 – Lecture 10

Data Mining and Predictive Analytics

Week 3

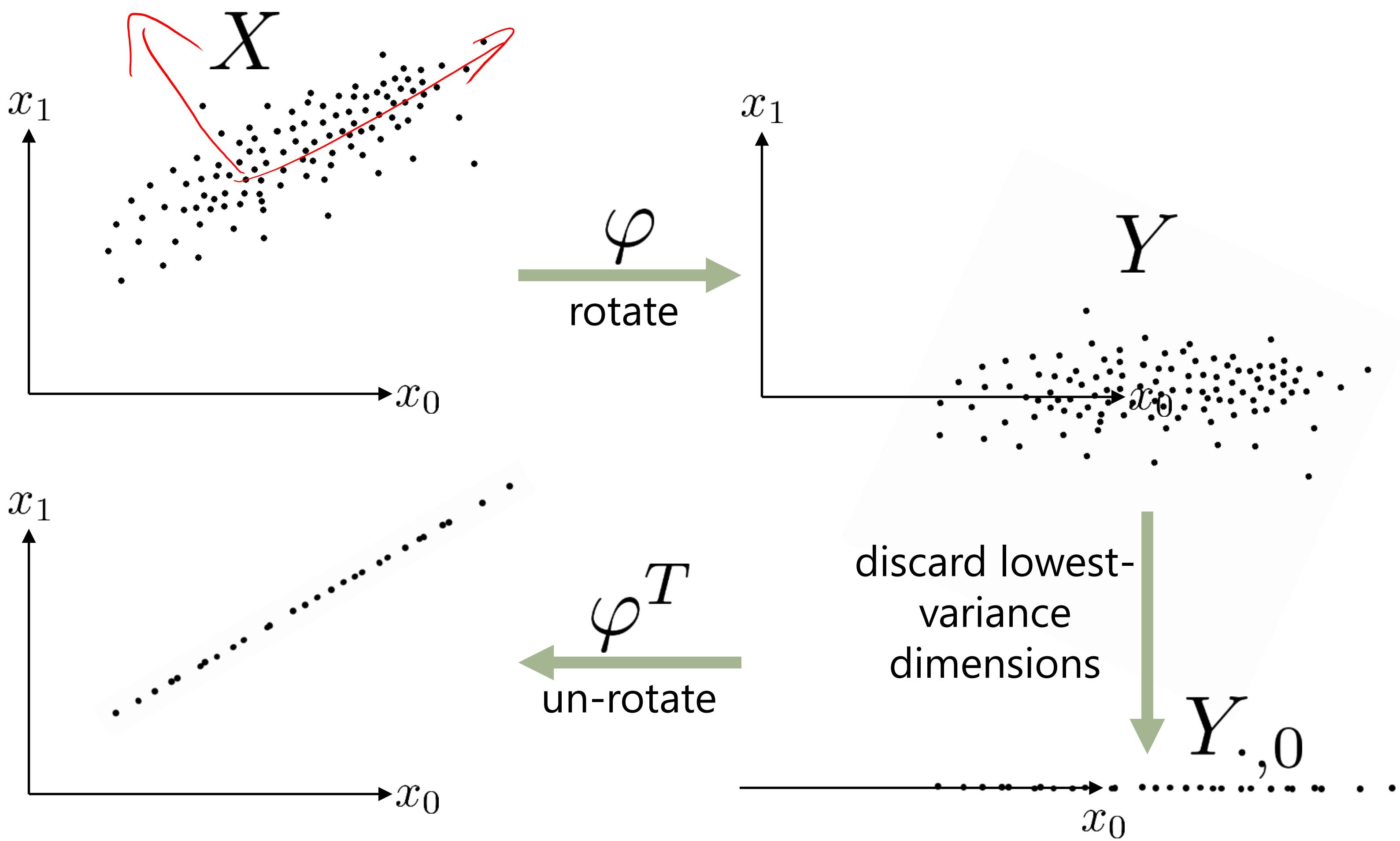
Why dimensionality reduction?

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)

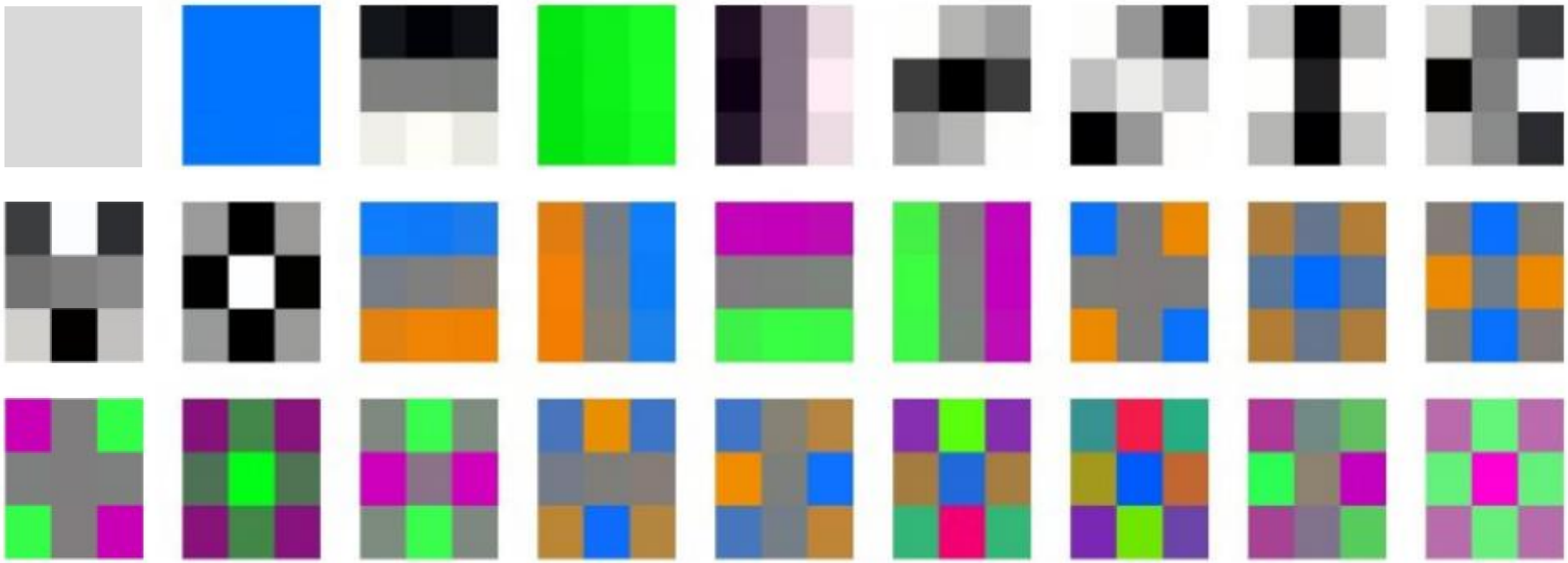
Principal Component Analysis



Principal Component Analysis

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

Color:



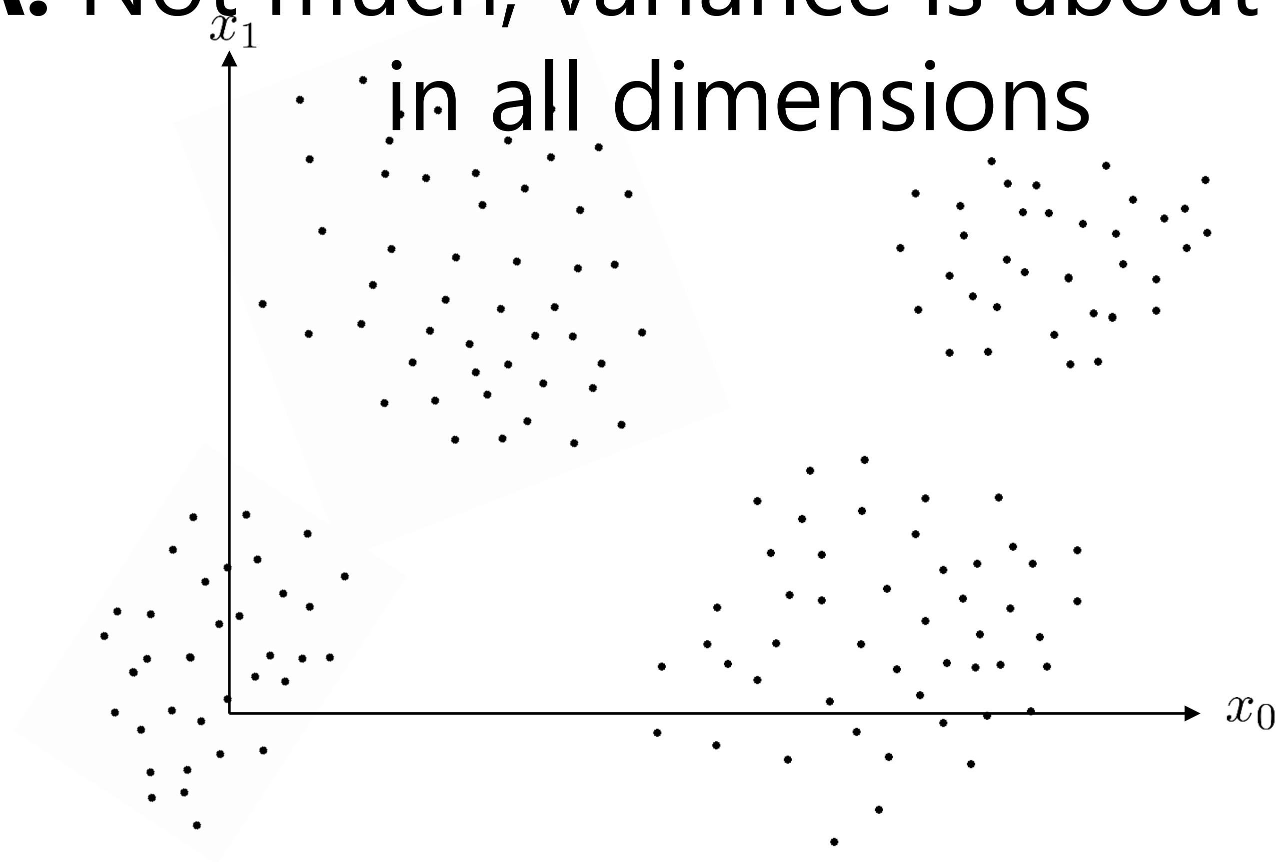
Principal Component Analysis

- 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

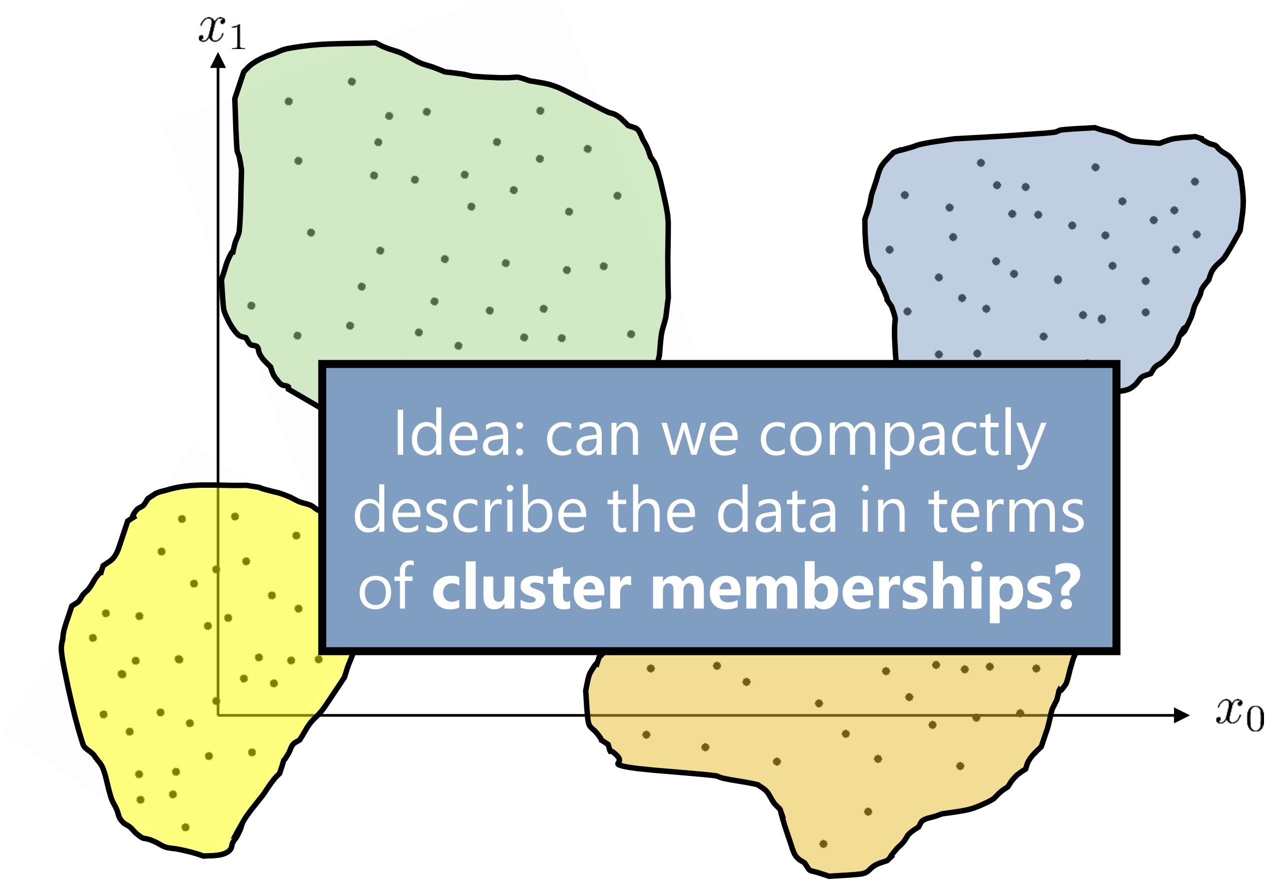
Q: What would PCA do with this data?

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



Clustering

But: The data are highly **clustered**



Idea: can we compactly describe the data in terms of **cluster memberships?**

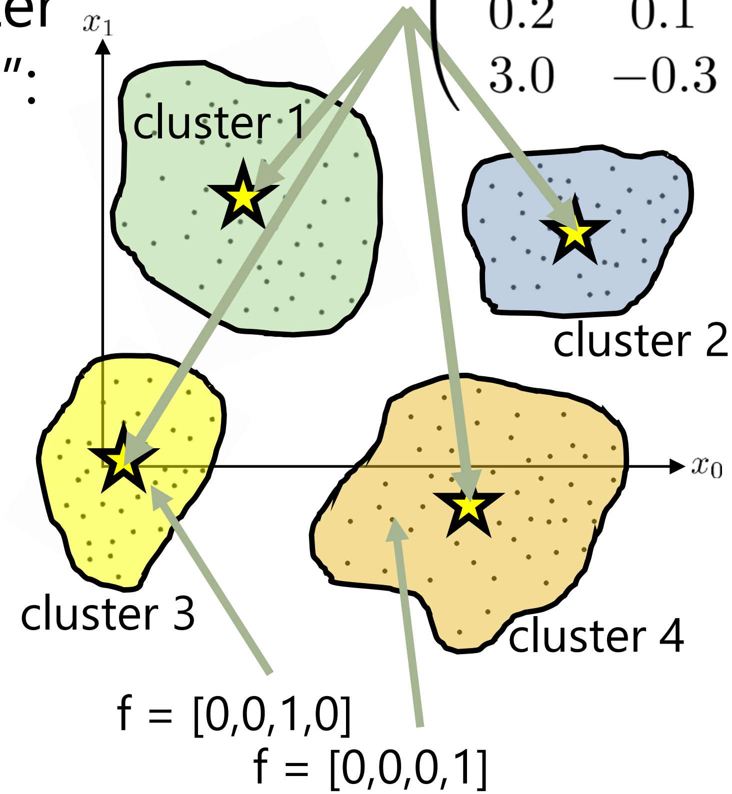
K-means Clustering

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":

$$\text{centroids} = \begin{pmatrix} 1.1 & 2.1 \\ 3.5 & 1.8 \\ 0.2 & 0.1 \\ 3.0 & -0.3 \end{pmatrix}$$



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)
```

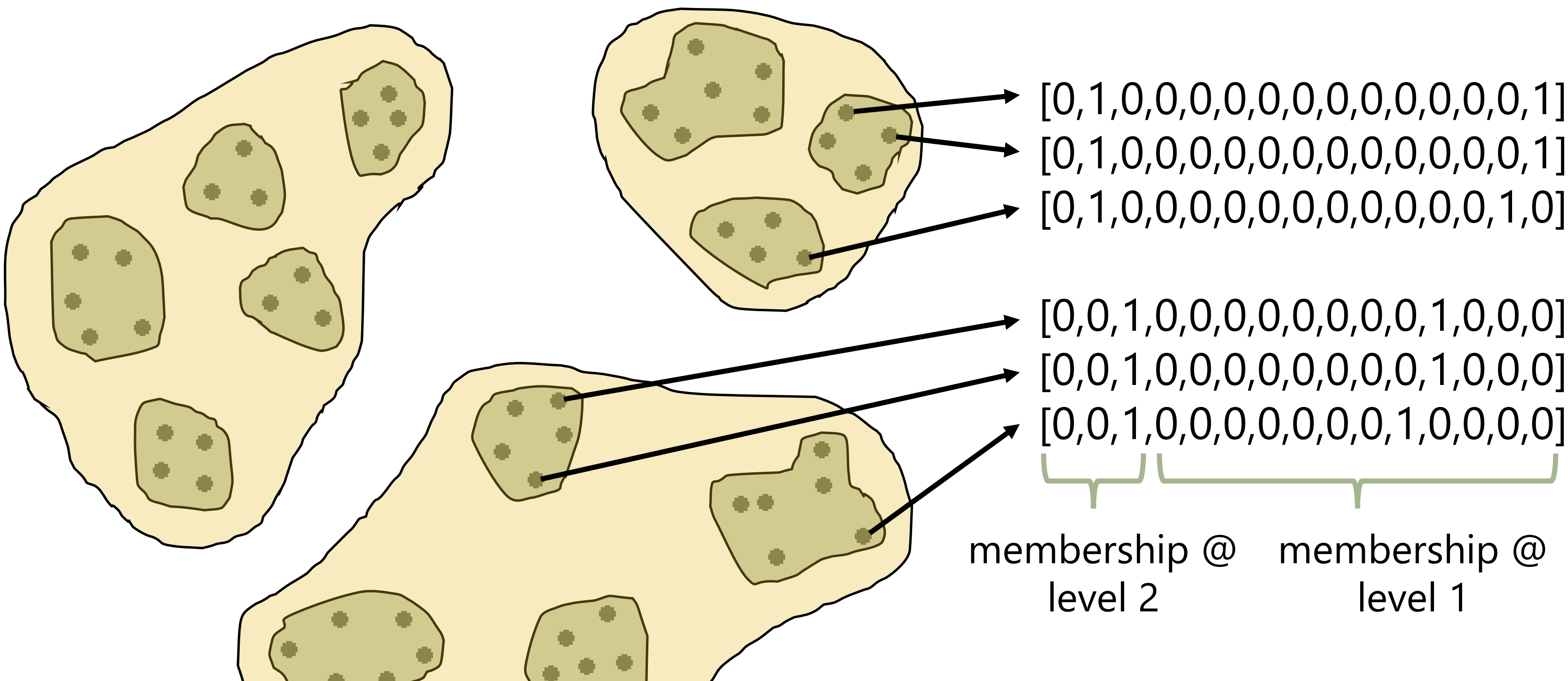
$$y_i = \arg\min_k \|X_i - C_k\|_2^2$$

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

(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

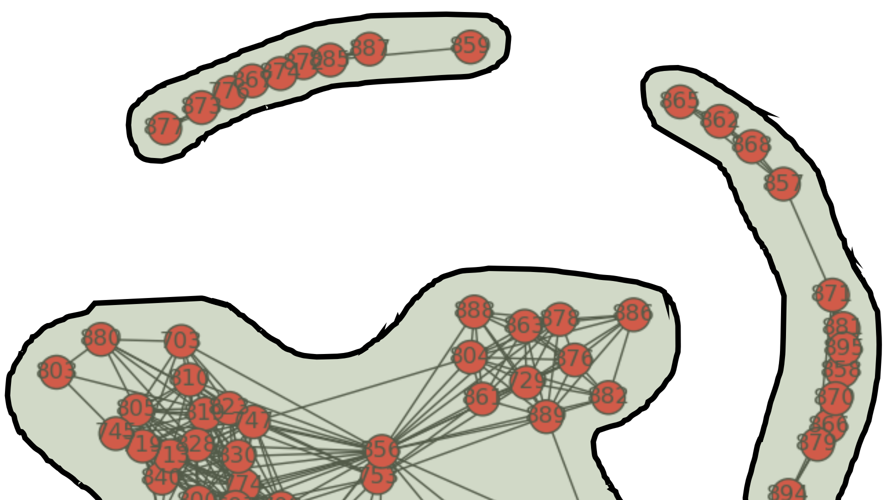
Hierarchical clustering

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
```

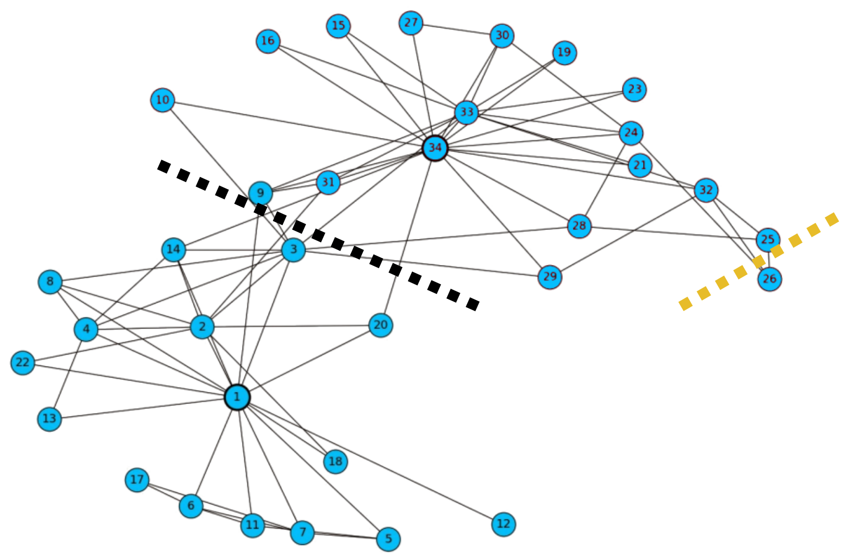
1. Connected components

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

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



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

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

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

Midterm on Monday!

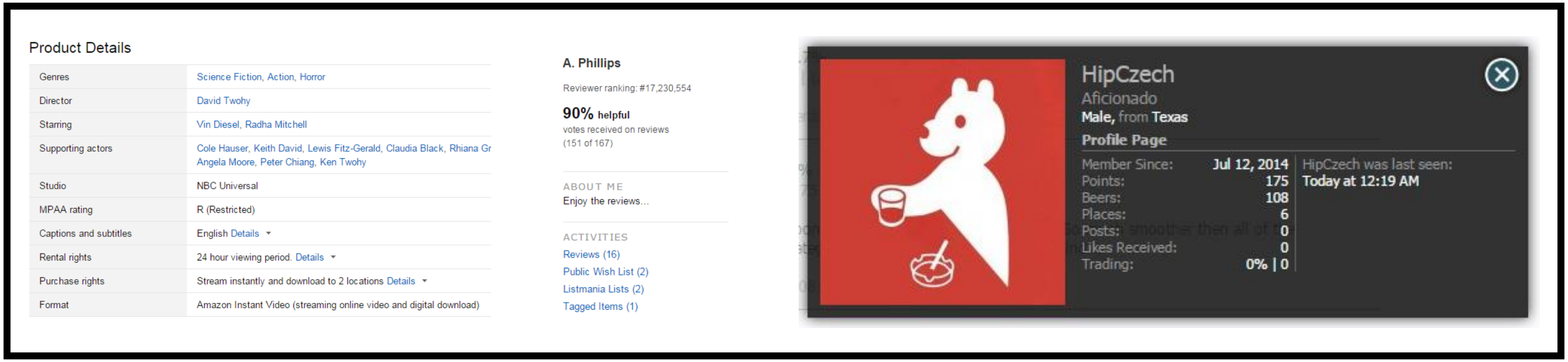
- Similar in format to last quarter's midterm
- Somewhat harder since you have a practice exam to work from (and it was just too easy last time!)
- Worth a little bit less (30% vs 25%)

CSE 190 – Lecture 10

Data Mining and Predictive Analytics

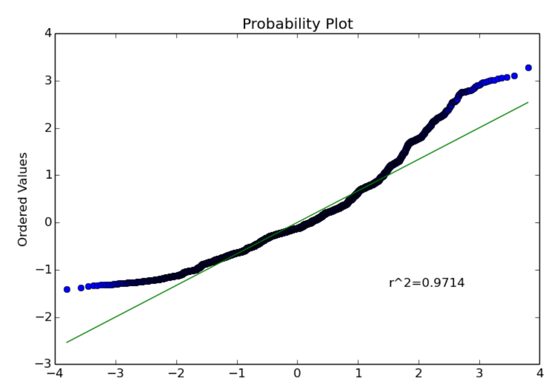
Midterm

1) Regression



How can we use **features** such as product properties and user demographics to make predictions about **real-valued** outcomes (e.g. star ratings)?

How can we prevent our models from **overfitting** by favouring simpler models over more complex ones?



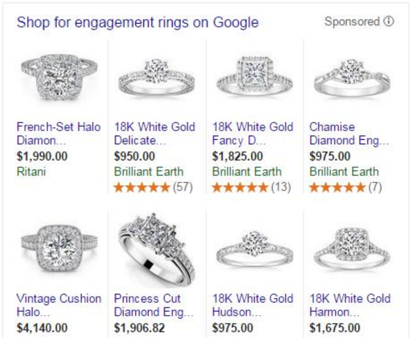
How can we assess our decision to optimize a particular error measure, like the MSE?

So far: Classification

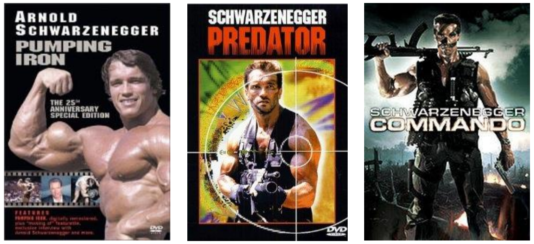
Next we adapted these ideas to **binary** or **multiclass** outputs

What animal is in this image?

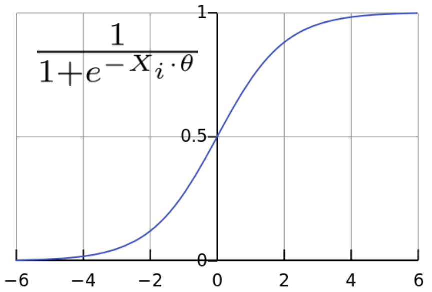
Will I **purchase** this product?



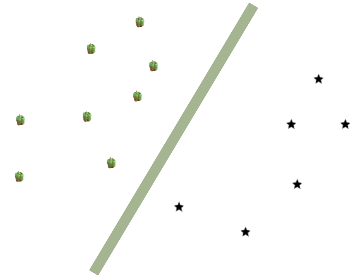
Will I **click on** this ad?



Combining features using naïve Bayes models

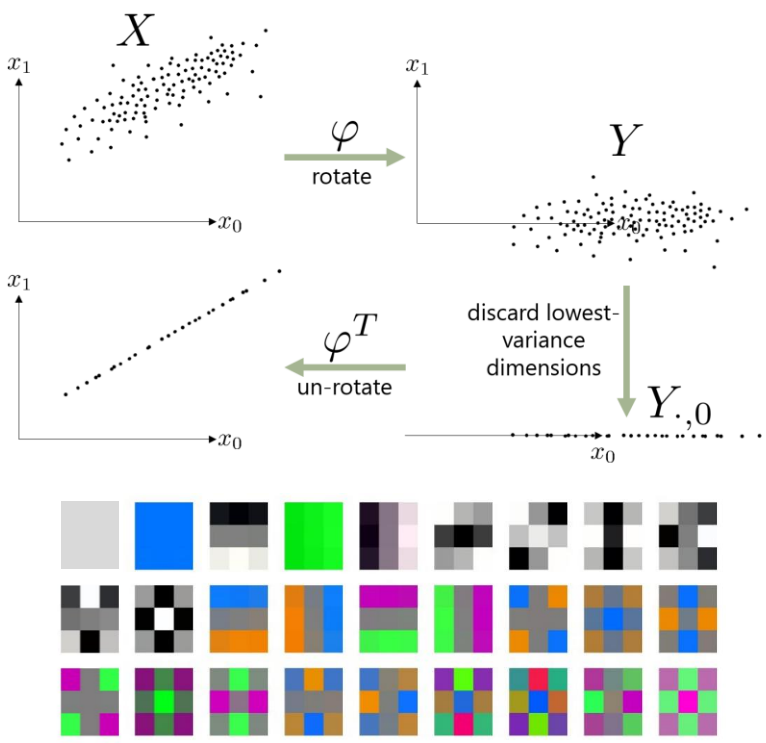


Logistic regression

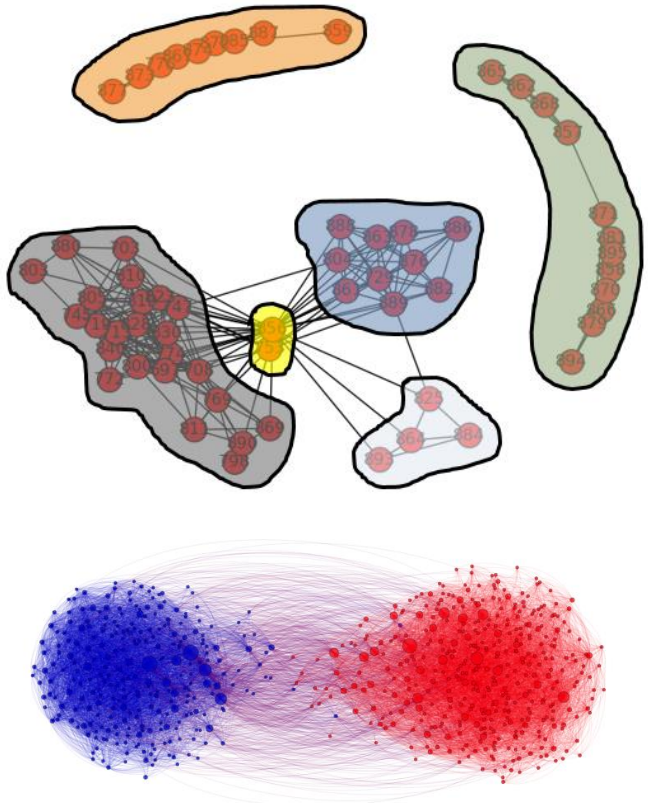


Support vector machines

Week 3: Dimensionality Reduction



Principal component analysis & clustering



Community detection

CSE 190 – Lecture 10

Data Mining and Predictive Analytics

Last quarter's midterm

Last quarter's 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

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

Last quarter's 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: 4.5

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

$$\left(0 + \frac{1}{25} + 0 + \frac{1}{25} + 0\right) / 5 \qquad \frac{1}{10}$$

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:

$$\begin{bmatrix} 4.5 \\ 5.0 \\ 4.5 \\ 4.0 \\ 4.5 \end{bmatrix} \simeq \begin{bmatrix} 1 & 1 & 0 & 3 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 2 \\ 1 & 0 & 1 & 2 \\ 1 & 1 & 1 & 2 \end{bmatrix} \theta$$

(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: y features: X parameters: θ

Last quarter's 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: what rat. would be, if all features 0

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

A: price goes up by 1$, rating goes down by 1

Last quarter's 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:

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

θ_3[price] + | A: |

Bre	Bru	L	D
1	0	0	0
0	0	1	0
0	1	0	0
0	0	1	1

2

and by completing the feature matrix using your representation:

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

(2 marks)

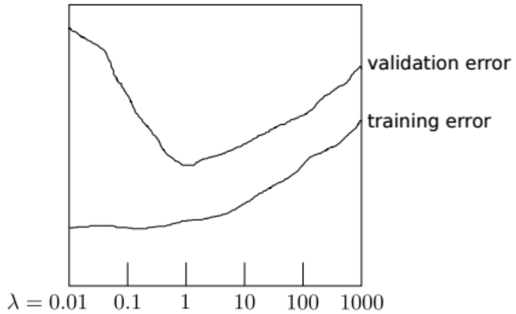
Last quarter's midterm

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:

Last quarter's 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	? > 23
model X, $\lambda = 10$	? > 23	? > 23
model Y, $\lambda = 1$	?	18.32
model Y, $\lambda = 10$	25.98	? > 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:

Last quarter's 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:	5	2	3	0

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:

$$\frac{5}{5+0} = 1 \qquad \frac{2}{2+3} = \frac{2}{5}$$

$$BER = \frac{1}{2}\left(1 - \frac{2}{5}\right)$$

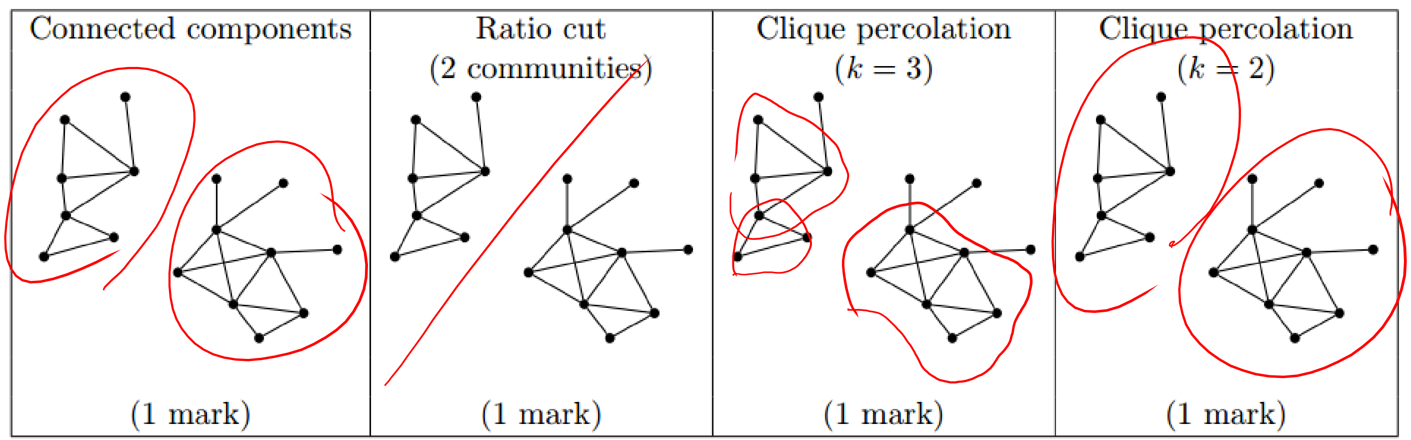
Last quarter's midterm

Section 3: Communities & clustering

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

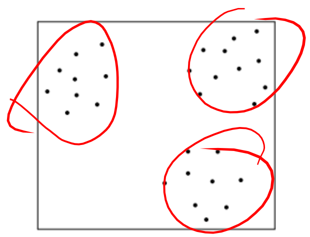
Recall three algorithms we saw in class to detect communities: connected components, ratio cut, and clique percolation (pseudocode is given as Algorithms 1, 2, and 3 at the end of the test).

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):



Last quarter's 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):

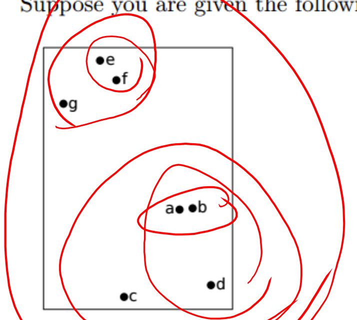


19. By replacing each point with one of the three centroids above, we have effectively 'compressed' the data, since each (2-d) point is replaced by a (1-d) integer. Another way to compress the data would be to perform Principal Component Analysis, and discard the lowest variance dimension, which would 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 190 – Lecture 10

Data Mining and Predictive Analytics

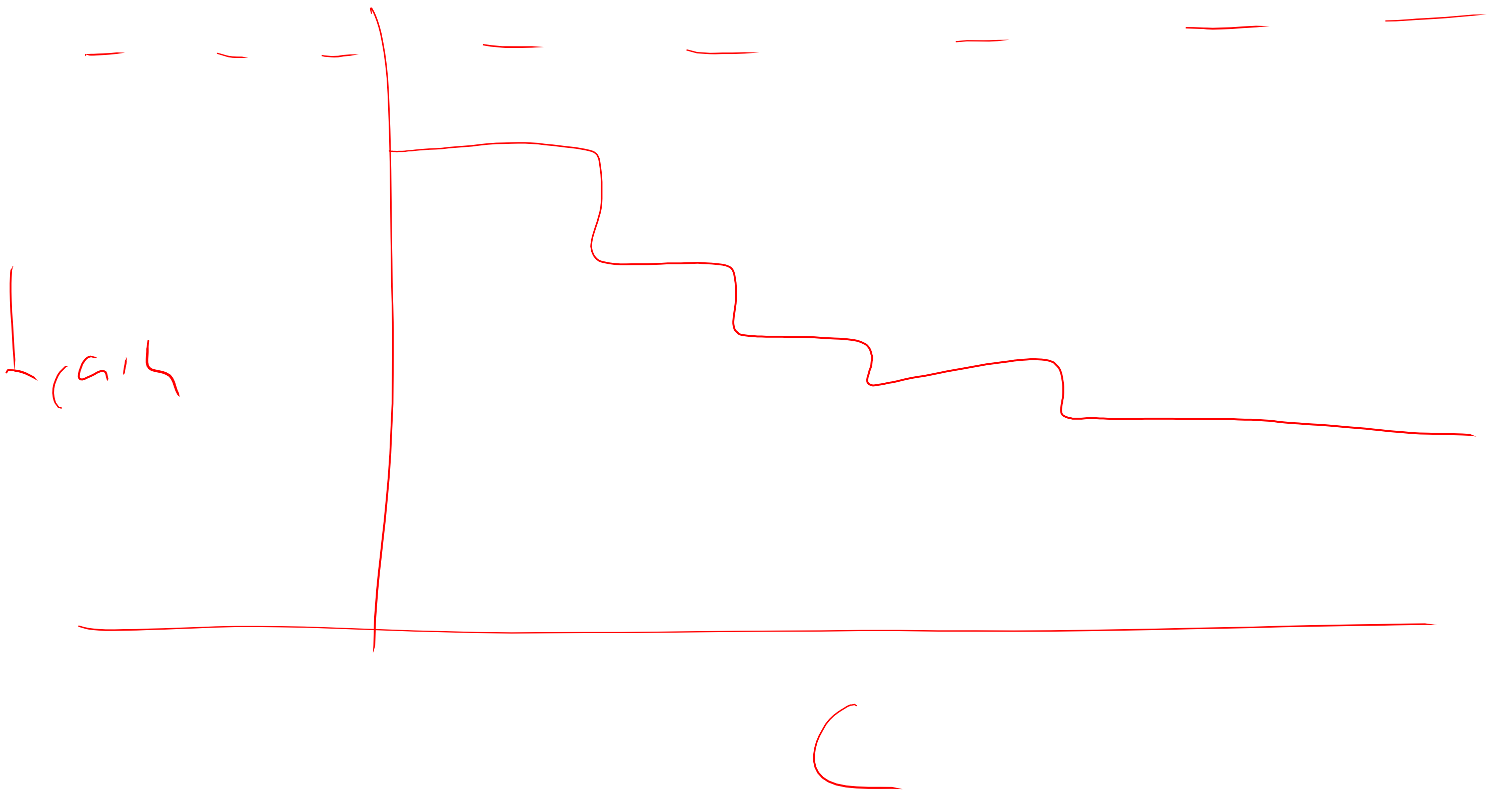
HW Questions

No reduction after degree 1 (HW1/wk1)

train = X[:2500]

test = X[2500:]

Train vs. lambda (Classification, HW1/wk2)



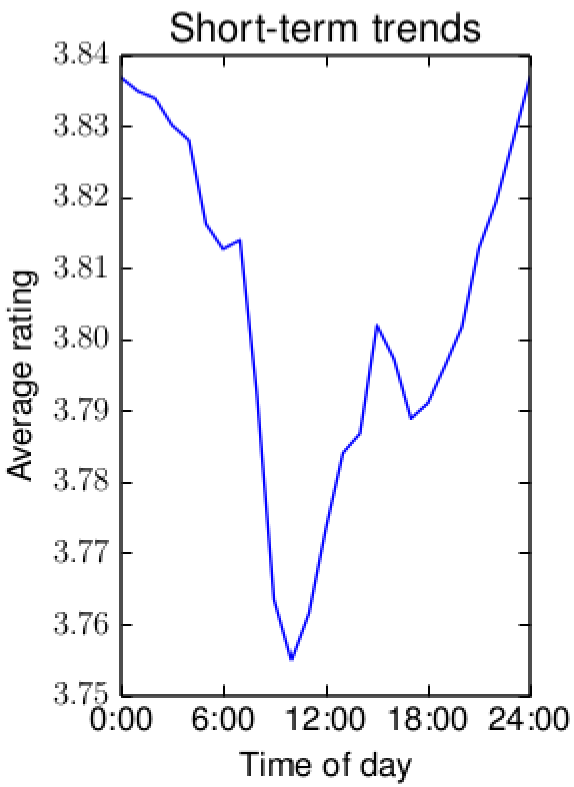
CSE 190 – Lecture 10

Data Mining and Predictive Analytics

Misc. questions

Representing the **day** as a feature

How would you build a feature to represent the time of **day?**



Representing the **day** as a feature

How would you build a feature to represent the time of **day?**

Interpretation of linear models

- Suppose we have a linear regression model to predict college GPA
- One of the features of this model encodes whether a student owns a car
- The fitted model looks like:

$$y = \ldots - 0.4[\text{owns a car}] + \ldots$$

Conclusion: "The GPA of the average student *who owns a car* is 0.4 lower than that of the average student"

Q: is this conclusion reasonable?