CSE 258 – Lecture 2
Web Mining and Recommender Systems

Supervised learning – Regression
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.

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.
Regression is one of the simplest supervised learning approaches to learn relationships between input variables (features) and output variables (predictions)
Linear regression assumes a predictor of the form

\[ y_i = x_i \cdot \theta \]

where \( X \theta = y \)

- **Matrix of features** (data)
- **Unknowns** (which features are relevant)
- **Vector of outputs** (labels)

(or \( Ax = b \) if you prefer)
Linear regression assumes a predictor of the form

\[ X \theta = y \]

Q: Solve for theta
A: \[ \theta = (X^T X)^{-1} X^T y \]
Example 1

How do preferences toward certain beers vary with age?
Example 1

**Beers:**

**Ratings/reviews:**

4.35/5 (row: 5.2%)

- Look: 4
- Smell: 4.25
- Taste: 4.5
- Feel: 4.25
- Overall: 4.25

Serving: 355 mL bottle poured into a 9 oz Libbey Embassy snifter (*bottled on: 08/15/14 11109*).

Appearance: Deep, dark near-black brown. Hazy, light brown fringes of foam and limited lacing; no head.

Smell: Roasted malt, vanilla, and some warming alcohol.

Taste: Roasted malts, cocoa, burnt caramel, molasses, vanilla and dark fruit. Bourbon barrel is hinted at but never takes over.

Mouthfeel: Medium to full body and light carbonation with a very lush, silky smooth feel.

Overall: Not as complex or intense as some newer barrel-aged stouts, but so smooth and balanced with all the elements tightly integrated.

*HipCzech, Yesterday at 05:38 AM*

**User profiles:**

- **HipCzech**
  - Affiliation: Afinazonado
  - Male, from Texas
  - **Profile Page**
    - Member Since: Jul 12, 2014
    - Today at 12:19 AM
    - Points: 175
    - Beers: 106
    - Places: 6
    - Posts: 0
    - Likes Received: 0
    - Trading: 0%
Example 1

50,000 reviews are available on
http://jmcauley.ucsd.edu/cse258/data/beer/beer_50000.json
(see course webpage)

See also – non-alcoholic beers:
http://jmcauley.ucsd.edu/cse258/data/beer/non-alcoholic-beer.json
Real-valued features

How do preferences toward certain beers vary with age?
How about $\text{ABV}$?

$$y_i = \theta_0 + \theta_1 \text{ age}$$

(code for all examples is on http://jmcauley.ucsd.edu/cse258/code/week1.py)
Example 1

Preferences vs \textbf{ABV}

\[
\begin{align*}
\text{ABV} & = \Theta_0 + \Theta_1 \text{ABV} + \Theta_2 \text{ABV}^2 \\
& \left[ \Theta_0, \Theta_1, \Theta_2 \right] \cdot \left[ 1, \text{ABV}, \text{ABV}^2 \right] \\
\text{error} & = \sum \left( y_i - x_i \cdot \Theta \right)^2
\end{align*}
\]
Example 2

Categorical features

\[ \theta \cdot x \]

How do beer preferences vary as a function of **gender**?

\[
\text{rating} = \theta_0 + \theta_1 \left[ \text{if male} \right] + \theta_2 \left[ \text{if female} \right]
\]

\[
x = \begin{cases} 
1, & \text{if male} \\
0, & \text{if female} 
\end{cases}
\]

Times females

(codes for all examples are on [http://jmcauley.ucsd.edu/cse258/code/week1.py](http://jmcauley.ucsd.edu/cse258/code/week1.py))
Linearly dependent features

\[ X = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \end{bmatrix} \]

\[ X^T X = \begin{bmatrix} 5 & 2 & 2 \\ 2 & 3 & 0 \\ 2 & 0 & 2 \\ a & b & 0 \end{bmatrix} \]

\[ \text{not invertible} \]

\[ \text{rating} = 3 + 1 (\text{if female}) + 2 (\text{male}) \]

\[ = 100 - 96 (\text{if female}) - 95 (\text{if male}) \]
Linearly dependent features

crating = \theta_0 + \theta_1 \text{ (int female)}

\begin{align*}
\text{Male:} & \quad c_{\theta_0} \\
\text{Female:} & \quad c_{\theta_0 + \theta_1}
\end{align*}
Exercise

How would you build a feature to represent the month, and the impact it has on people’s rating behavior?

crating = \theta_0 + \theta_1 [\text{month}]

January = 1
February = 2
March = 3
\[ x_i = \Theta_0 + \Theta_1 [\text{if } j = 1] + \Theta_2 [\text{if } j = 2] + \ldots + \Theta_n [\text{if } j = n] \]

\[ x_i = [1, 0, 0, 0, 1, 0, \ldots] \]
What does the data actually look like?

Season vs. rating (overall)
Example 3

Random features

What happens as we add more and more random features?

(code for all examples is on http://jmcauley.ucsd.edu/cse258/code/week1.py)
CSE 258 – Lecture 2
Web Mining and Recommender Systems

Regression Diagnostics
Today: Regression diagnostics

**Mean-squared error (MSE)**

\[
\frac{1}{N} \| y - X\theta \|^2_2 \\
= \frac{1}{N} \sum_{i=1}^{N} (y_i - X_i \cdot \theta)^2
\]
Q: Why MSE (and not mean-absolute-error or something else)
Regression diagnostics

Assume: small errors are common;
large errors are very uncommon.

\[ y_i - x_i \cdot \theta \]

\[ |\delta_i| = \text{prediction + error} \]

\[ y_i = x_i \cdot \theta + \mathcal{N}(0, \sigma) \]
Regression diagnostics

\[ P_0(y \mid x) = \prod_{i} \frac{1}{\sqrt{2\pi \sigma}} \exp\left( -\frac{(y_i - x_i \cdot \theta)^2}{2\sigma^2} \right) \]

\[ \max_{\theta} P_0(y \mid x) = \max_{\theta} \prod_{i} \exp\left( -\frac{(y_i - x_i \cdot \theta)^2}{2\sigma^2} \right) \]

\[ \theta \in \mathbb{R}^{m \times 1} \]
Coefficient of determination

Q: How low does the MSE have to be before it’s “low enough”?
A: It depends! The MSE is proportional to the variance of the data.
Coefficient of determination
(R^2 statistic)

Mean:  \[ \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \]

Variance:  \[ \text{Var}(\hat{y}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y})^2 \]

MSE:  \[ \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i \cdot \hat{\theta})^2 \]
Regression diagnostics

Coefficient of determination
(R^2 statistic)

\[ FVU(f) = \frac{MSE(f)}{Var(y)} \]

(FVU = fraction of variance unexplained)

\[
\begin{align*}
FVU(f) &= 1 \quad \text{Trivial predictor} \\
FVU(f) &= 0 \quad \text{Perfect predictor}
\end{align*}
\]
Regression diagnostics

Coefficient of determination
\((R^2\ \text{statistic})\)

\[ R^2 = 1 - FVU(f) = 1 - \frac{MSE(f)}{\text{Var}(y)} \]

\(R^2 = 0\) \quad \text{Trivial predictor}
\(R^2 = 1\) \quad \text{Perfect predictor}
Q: But can’t we get an $R^2$ of 1 (MSE of 0) just by throwing in enough random features?

A: Yes! This is why MSE and $R^2$ should always be evaluated on data that wasn’t used to train the model.

A good model is one that generalizes to new data.
Overfitting

When a model performs well on **training** data but doesn’t generalize, we are said to be **overfitting**
Overfitting

When a model performs well on training data but doesn’t generalize, we are said to be overfitting.

Q: What can be done to avoid overfitting?
Occam’s razor

“Among competing hypotheses, the one with the fewest assumptions should be selected”
Occam’s razor

\[ X \theta = y \]

“hypothesis”

**Q:** What is a “complex” versus a “simple” hypothesis?
$$r_{avg} = O_0 + O_1 ANV + O_2 ANV^2 + O_3 \ldots \ldots$$

$$O(2) \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \q
A1: A “simple” model is one where theta has few non-zero parameters (only a few features are relevant)

A2: A “simple” model is one where theta is almost uniform (few features are significantly more relevant than others)
Occam’s razor

**A1:** A “simple” model is one where theta has few non-zero parameters

\[ \| \theta \|_1 \text{ is small} \]

**A2:** A “simple” model is one where theta is almost uniform

\[ \| \theta \|_2 \text{ is small} \]
Proof

\[ \text{height} = O(0) + O(1) \text{ age } + O(2) \text{ shoe size} \]

\[ \frac{a}{55} \]

\[ O(1) \]

\[ \frac{1}{5} \]

\[ \frac{O(2)}{O(2)} \]

\[ \| O(1) \|_2^2 > \frac{\| O(2) \|_2^2}{\| O(2) \|_1} \]

\[ \| O(1) \|_1 = \frac{\| O(2) \|_2}{\| O(2) \|_1} \]
Regularization is the process of penalizing model complexity during training.

\[ \text{arg min}_\theta = \frac{1}{N} \| y - X\theta \|_2^2 + \lambda \| \theta \|_2^2 \]

- **MSE**
- **(l2) model complexity**
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?
Optimizing the (regularized) model

\[ \arg \min_\theta = \frac{1}{N} \left\| y - X \theta \right\|^2_2 + \lambda \left\| \theta \right\|^2_2 \]

- Could look for a closed form solution as we did before
- Or, we can try to solve using \textit{gradient descent}
Gradient descent:

1. Initialize $\theta$ at random
2. While (not converged) do
   \[ \theta := \theta - \alpha f'(\theta) \]

All sorts of annoying issues:
- How to initialize theta?
- How to determine when the process has converged?
- How to set the step size alpha

These aren’t really the point of this class though
Optimizing the (regularized) model

\[ f(\theta) = \frac{1}{N} \| y - X\theta \|_2^2 + \lambda \| \theta \|_2^2 \]

\[ \frac{\partial f}{\partial \theta_k} = \frac{1}{N} \sum_i (y_i - x_i \cdot \theta)^2 + \lambda \sum_j \theta_j^2 \]

\[ = \frac{1}{N} \sum_i 2 \theta_k (y_i - x_i \cdot \theta) + \lambda \theta_k \]

\[ - 2 x_{ik} \]
Optimizing the (regularized) model

Gradient descent in scipy:

(code for all examples is on http://jmcauley.ucsd.edu/cse258/code/week1.py)

(see “ridge regression” in the “sklearn” module)
Model selection

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

How much should we trade-off accuracy versus complexity?

Each value of lambda generates a different model. \textbf{Q:} How do we select which one is the best?
Model selection

How to select which model is best?

A1: The one with the lowest training error?

A2: The one with the lowest test error?

We need a third sample of the data that is not used for training or testing
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
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.
Model selection
Summary of Week 1: Regression

• Linear regression and least-squares
  • (a little bit of) feature design
• Overfitting and regularization
  • Gradient descent
• Training, validation, and testing
  • Model selection
Coming up!

An exciting case study (i.e., my own research)!

This photo recently won the Andrews award for the 'most perfect timing of a Nature photograph', I can see why.
submitted 29 days ago by SICK_OF_ to /r/pics
11 points
1 comment

Perfect moment bird (ex-post from /r/pics)
submitted 25 days ago by 123imAwesome to /r/photoshopbattles
35 points
11 comments

Bird shot at the perfect moment
submitted 25 days ago by arbili to /r/pics
2712 points
166 comments

Perfect timing.
submitted 4 months ago by animalpath to /r/pics
2555 points
718 comments

Perfect timing.
submitted 2 months ago by presaging to /r/aww
12 points
1 comment

NOM! (Photo by: Bohemian Waxwing)
submitted 2 months ago by feveritehike [deleted] to /r/PerfectTiming
1117 points
11 comments

A bohemian waxwing eating a berry
submitted 4 months ago by HazeySynth to /r/pics
39 points
1 comment

Timing is Everything
submitted 5 months ago by Xnico378X to /r/pics
10 points
1 comment
Homework is **available** on the course webpage

http://cseweb.ucsd.edu/classes/fa17/cse258-a/files/homework1.pdf

Please submit it by the beginning of the **week 3** lecture (Oct 16)

All submissions should be made as **pdf** files on gradescope
Questions?