CSE 258 – Lecture 1.5
Web Mining and Recommender Systems

Supervised learning – Regression
Supervised learning is the process of trying to infer from labeled data the underlying function that produced the labels associated with the data.
Given **labeled training data** of the form

\[\{(data_1, label_1), \ldots, (data_n, label_n)\}\]

Infer the function

\[f(data) \rightarrow \text{labels}\]
Example

Suppose we want to build a movie recommender

e.g. which of these films will I rate highest?
**Q:** What are the labels?

**A:** ratings that others have given to each movie, and that I have given to other movies.
Q: What is the data?

A: features about the movie and the users who evaluated it

Movie features: genre, actors, rating, length, etc.

User features: age, gender, location, etc.
Movie recommendation:
\[ f(\text{data}) \rightarrow \text{labels} \]
\[ = \]
\[ f(\text{user features, movie features}) \rightarrow \text{star rating} \]
Design a system based on prior knowledge, e.g.

```python
def prediction(user, movie):
    if (user['age'] <= 14):
        if (movie['mpaa_rating']) == "G"):
            return 5.0
        else:
            return 1.0
    else if (user['age'] <= 18):
        if (movie['mpaa_rating']) == "PG"):
            return 5.0
    .... Etc.
```

Is this supervised learning?
Solution 2

Identify words that I frequently mention in my social media posts, and recommend movies whose plot synopses use similar types of language.

Is this supervised learning?

argmax similarity(synopsis, post)
Identify which attributes (e.g. actors, genres) are associated with positive ratings. Recommend movies that exhibit those attributes.

Is this *supervised learning*?
Solution 1

(design a system based on prior knowledge)

Disadvantages:
- Depends on possibly false assumptions about how users relate to items
- Cannot adapt to new data/information

Advantages:
- Requires no data!
Solution 2

(identify similarity between wall posts and synopses)

Disadvantages:
• Depends on possibly false **assumptions** about how users relate to items
• May not be adaptable to new settings

Advantages:
• Requires data, but does not require **labeled** data
Solution 3

(identify attributes that are associated with positive ratings)

Disadvantages:
• Requires a (possibly large) dataset of movies with labeled ratings

Advantages:
• Directly optimizes a measure we care about (predicting ratings)
• Easy to adapt to new settings and data
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.

**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

\[ X\theta = y \]

(or \( Ax = b \) if you prefer)

matrix of features (data)

unknowns (which features are relevant)

vector of outputs (labels)
Motivation: height vs. weight

Q: Can we find a line that (approximately) fits the data?
Motivation: height vs. weight

Q: Can we find a line that (approximately) fits the data?

• If we can find such a line, we can use it to make predictions (i.e., estimate a person's weight given their height).
• How do we formulate the problem of finding a line?
• If no line will fit the data exactly, how to approximate?
  • What is the "best" line?
Recap: equation for a line

What is the formula describing the line?

$y = mx + b$

Weight = $m \times$ Height + $b$
Recap: equation for a line

What about in more dimensions?

Weight = \( m_1 \times \text{Height} + m_2 \times \text{age} + b \)

\[ y = m_1 x_1 + m_2 x_2 + b \]
Recap: equation for a line as an inner product

What about in more dimensions?

Weight = (Height, age, 1) \cdot (m_1, m_2, b)

y = (x_1, x_2, 1) \cdot (m_1, m_2, b)
\[ y = X \theta \]
\[ y_i = \begin{bmatrix} 180 & 165 \end{bmatrix} \]
\[ x_i = 0 \]
\[ \begin{bmatrix} 170 & 0 & 22 & 1 \end{bmatrix} \]
Linear regression assumes a predictor of the form

\[ X\theta = y \]

**Q:** Solve for theta

**A:**

\[ \theta = (X^TX)^{-1}X^Ty \]
Example 1

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

**Beer advocate**

**Beers:**

**Ratings/reviews:**

4.35/5 5.2%

Serve: 355 ml bottle poured into a 9 oz Libbey Embassy snifter (bottled on 08/24/11). Appearance: Deep, dark near-black brown. Hazy, light brown fringe 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 smooth and balanced with all the elements tightly integrated.

**User profiles:**

**HipCzech**

Alicacians

Male, from Texas

Member Since: Jul 12, 2014

Profile Page:

- Points: 175
- Beers: 108
- Places: 6
- Posts: 0
- Likes Received: 0
- Trading: 0%

Today at 12:19 AM
Example 1

50,000 reviews are available on
http://jmcauley.ucsd.edu/cse258/data/beer/beer_50000.json
(see course webpage)
Example 1

Real-valued features

How do preferences toward certain beers vary with age?
How about ABV?

\[
\text{rating} = \theta_0 + \theta_1 \times \text{age} \\
[\theta_0, \theta_1] \cdot [1, \text{age}]
\]

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

What is the interpretation of:

$$\theta = \left( 3.4, 10e^{-7} \right)$$

(code for all examples is on http://jmcauley.ucsd.edu/cse258/code/week1.py)
Example 2

Categorical features

How do beer preferences vary as a function of gender?

(code for all examples is on http://jmcauley.ucsd.edu/cse258/code/week1.py)
E.g. How does rating vary with gender?
Example 2

\( \theta_0 \) is the (predicted/average) rating for males

\( \theta_1 \) is the **how much higher** females rate than males (in this case a negative number)

We’re really still fitting a line though!
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)
Exercise

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