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

\[ X\theta = y \]

(matrix of features (data) \rightarrow unknowns (which features are relevant) \rightarrow 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:** \[ \theta = (X^T X)^{-1} X^T y \]
How do preferences toward certain beers vary with age?
Example 1

**Beers:**

![Image of a beer bottle](image)

**Ratings/reviews:**

- **BA SCORE:** 100 world-class
- **THE BRU5:** 95 world-class
- **Brewed by:** Goose Island Beer Co., Illinois, United States
- **Style:** American Double/Imperial Stout
- **ABV:** 13.86%
- **Availability:** Winter
- **Notes/Commercial Description:** 00 IBU

**User profiles:**

- **HipCzech:**
  - **Affiliation:** Aclicionado
  - **Male, from Texas
  - **Profile Page:**
    - **Member Since:** Jul 12, 2014
    - **Points:** 175
    - **Beers:** 108
    - **Places:** 6
    - **Posts:** 0
    - **Loves:** 0
    - **Trading:** 0% / 0
Example 1

50,000 reviews are available on http://jmcauley.ucsd.edu/cse158/data/beer/beer_50000.json (see course webpage)

See also – non-alcoholic beers: http://jmcauley.ucsd.edu/cse158/data/beer/non-alcoholic-beer.json
Example 1

Real-valued features

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

(code for all examples is on http://jmcauley.ucsd.edu/cse158/code/week1.py)
Example 1

Preferences vs ABV
Real-valued features

What is the interpretation of:

$$\theta = (3.4, 10e^{-7})$$
Categorical features

How do beer preferences vary as a function of gender?

(code for all examples is on http://jmcauley.ucsd.edu/cse158/code/week1.py)
Linearly dependent features
Exercise

How would you build a feature to represent the **month**, and the impact it has on people’s rating behavior?
What does the data actually look like?

Season vs. rating (overall)
CSE 158 – 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
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:

Variance:

MSE:
Regression diagnostics

Coefficient of determination
(R^2 statistic)

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

(FVU = fraction of variance unexplained)

\[ FVU(f) = 1 \quad \rightarrow \quad \text{Trivial predictor} \]
\[ FVU(f) = 0 \quad \rightarrow \quad \text{Perfect predictor} \]
Regression diagnostics

**Coefficient of determination**
(R^2 statistic)

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

\[ R^2 = 0 \quad \text{Trivial predictor} \]

\[ R^2 = 1 \quad \text{Perfect predictor} \]
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.
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”
Q: What is a “complex” versus a “simple” hypothesis?
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

A2: A “simple” model is one where theta is almost uniform
“Proof”
Regularization is the process of penalizing model complexity during training.

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

- 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

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

\[f(\theta)\]

- We no longer have a convenient closed-form solution for theta
- Need to resort to some form of approximation algorithm
Optimizing the (regularized) model

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

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

\[ \frac{\partial f}{\partial \theta_k} \quad ? \]
Optimizing the (regularized) model

Gradient descent in scipy:

(code for all examples is on http://jmcauley.ucsd.edu/cse158/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. 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
Homework

Homework is available on the course webpage


Please submit it at the beginning of the week 3 lecture (Jan 23)

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