

# CSE 255 – Lecture 2

Data Mining and Predictive Analytics

Supervised learning – Regression

# 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

**Regression** is one of the simplest supervised learning approaches to learn relationships between input variables (features) and output variables (predictions)

# 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)

# Linear regression

**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

## Beeradvocate

### Beers:



Displayed for educational use only; do not reuse.

BA SCORE <b>100</b> world-class 9,587 Ratings	THE BROS <b>95</b> world-class (view ratings)	Ratings: 9,587 Reviews: 2,537 rAvg: 4.59 pDev: 9.59% Wants: 2,109 Gots: 4,563   FT: 472
--	--	--

Brewed by:  
Goose Island Beer Co.   
Illinois, United States

Style | ABV  
American Double / Imperial Stout | 13.80% ABV

Availability: Winter

Notes/Commercial Description:  
60 IBU

(Beer added by: drewbage on 06-26-2003)

### Ratings/reviews:



**4.35/5** rDev -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: 08AUG14 1109").

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 so smooth and balanced with all the elements tightly integrated.

HipCzech, Yesterday at 05:38 AM

### User profiles:



<b>HipCzech</b>		
Aficionado		
Male, from Texas		
<b>Profile Page</b>		
Member Since:	<b>Jul 12, 2014</b>	HipCzech was last seen:
Points:	<b>175</b>	Today at 12:19 AM
Beers:	<b>108</b>	
Places:	<b>6</b>	
Posts:	<b>smoother than all of</b>	<b>0</b>
Likes Received:	<b>0</b>	
Trading:	<b>0%   0</b>	

# Example 1

50,000 reviews are available on

[http://jmcauley.ucsd.edu/cse255/data/beer/beer\\_50000.json](http://jmcauley.ucsd.edu/cse255/data/beer/beer_50000.json)

(see course webpage)

See also – non-alcoholic beers:

<http://jmcauley.ucsd.edu/cse255/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/cse255/code/week1.py>)

# Example 1

## Preferences vs **ABV**

# Example 1

## Real-valued features

What is the interpretation of:

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

## Example 2

### Categorical features

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

(code for all examples is on <http://jmcauley.ucsd.edu/cse255/code/week1.py>)

# Linearly dependent features

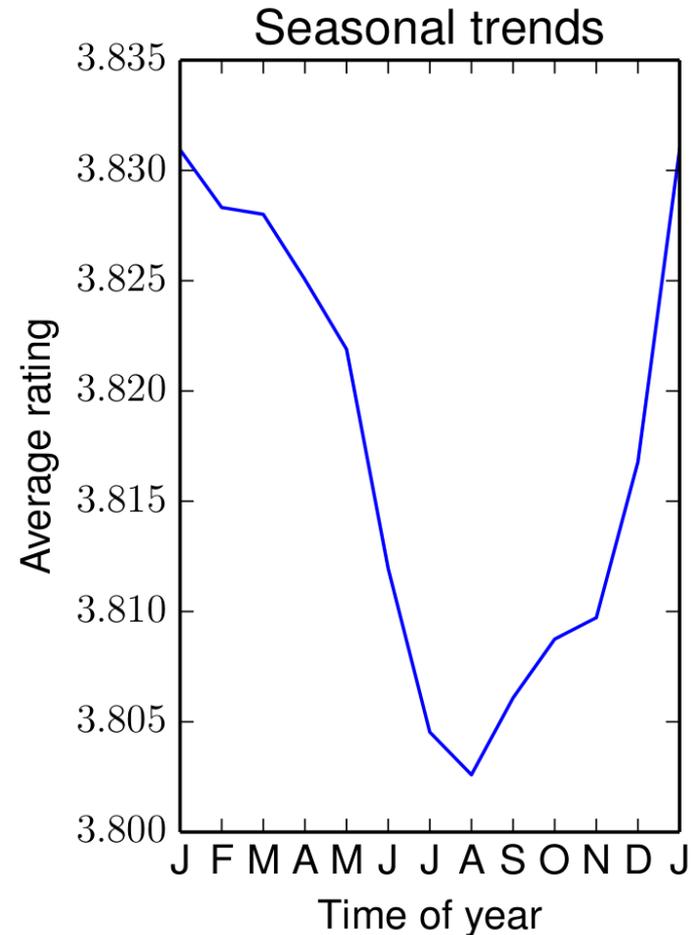
# Exercise

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

# Exercise

# 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/cse255/code/week1.py>)

# CSE 255 – Lecture 2

Data Mining and Predictive Analytics

Regression Diagnostics

# Today: Regression diagnostics

## **Mean-squared error (MSE)**

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

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

# Regression diagnostics

**Q:** Why MSE (and not mean-absolute-error or something else)

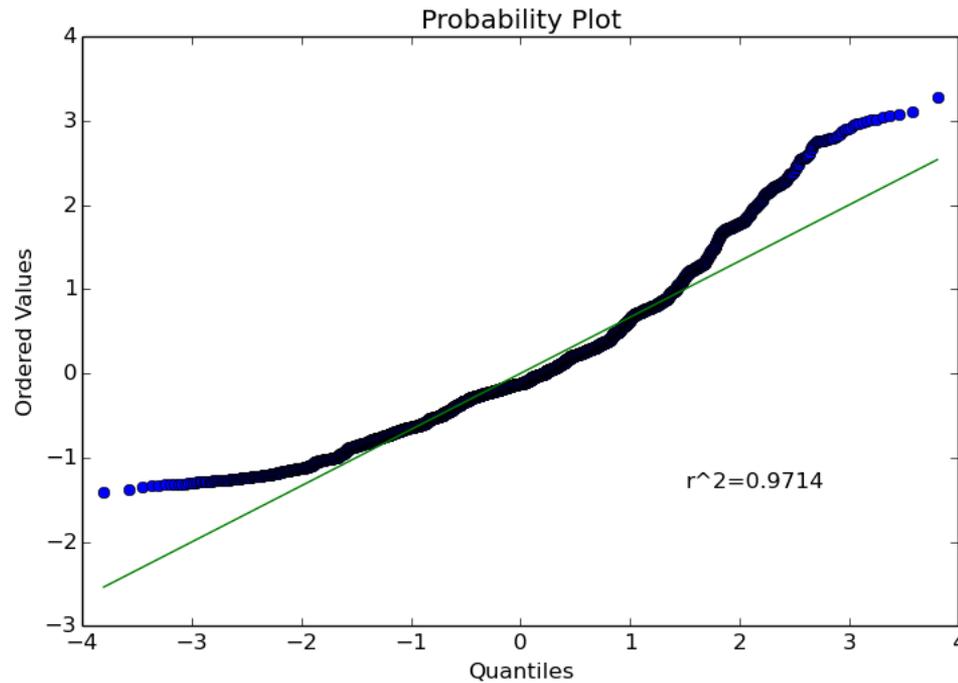
# Regression diagnostics

# Regression diagnostics

# Regression diagnostics

# Regression diagnostics

## Quantile-Quantile (QQ)-plot



## **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

# Regression diagnostics

## **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)

$FVU(f) = 1$   Trivial predictor

$FVU(f) = 0$   Perfect predictor

# Regression diagnostics

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

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

$R^2 = 0$   Trivial predictor

$R^2 = 1$   Perfect predictor

# Overfitting

**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**

**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?

# Occam's razor

**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  $\longrightarrow \|\theta\|_1$  is small

**A2:** A "simple" model is one where theta is almost uniform  $\longrightarrow \|\theta\|_2$  is small

"Proof"

# Regularization

**Regularization** is the process of penalizing model complexity during training

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

MSE



(l2) model complexity



# 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?

# Optimizing the (regularized) model

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

- 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

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

$$\frac{\partial f}{\partial \theta_k} \text{ ?}$$

# Optimizing the (regularized) model

## Gradient descent in scipy:

(code for all examples is on <http://jmcauley.ucsd.edu/cse255/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

# 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

# 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

# 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 one 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



NOM! (Photo by: Bohemian Waxwing)

submitted 2 months ago by favoritehell0 [deleted] to /r/PerfectTiming

1117 points  
1 comment



Perfect moment bird (ex-post from r/pics)

submitted 25 days ago by 123imAwesome to /r/photoshobbattles

36 points  
1 comment



A bohemian waxwing eating a berry

submitted 4 months ago by HazySynth to /r/pics

39 points  
1 comment



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  
71 comments



Perfect timing.

submitted 2 months ago by presaging to /r/aww

12 points  
1 comment



Timing is Everything

submitted 5 months ago by Xnicko378X to /r/pics

10 points  
1 comment

# Homework

Homework is **available** on the course  
webpage

[http://cseweb.ucsd.edu/classes/fa15/cse255-  
a/files/homework1.pdf](http://cseweb.ucsd.edu/classes/fa15/cse255-a/files/homework1.pdf)

Please submit it at the beginning of the  
**week 3** lecture (Oct 12)

Questions?