DATA 8

Spring 2020

Lecture 34

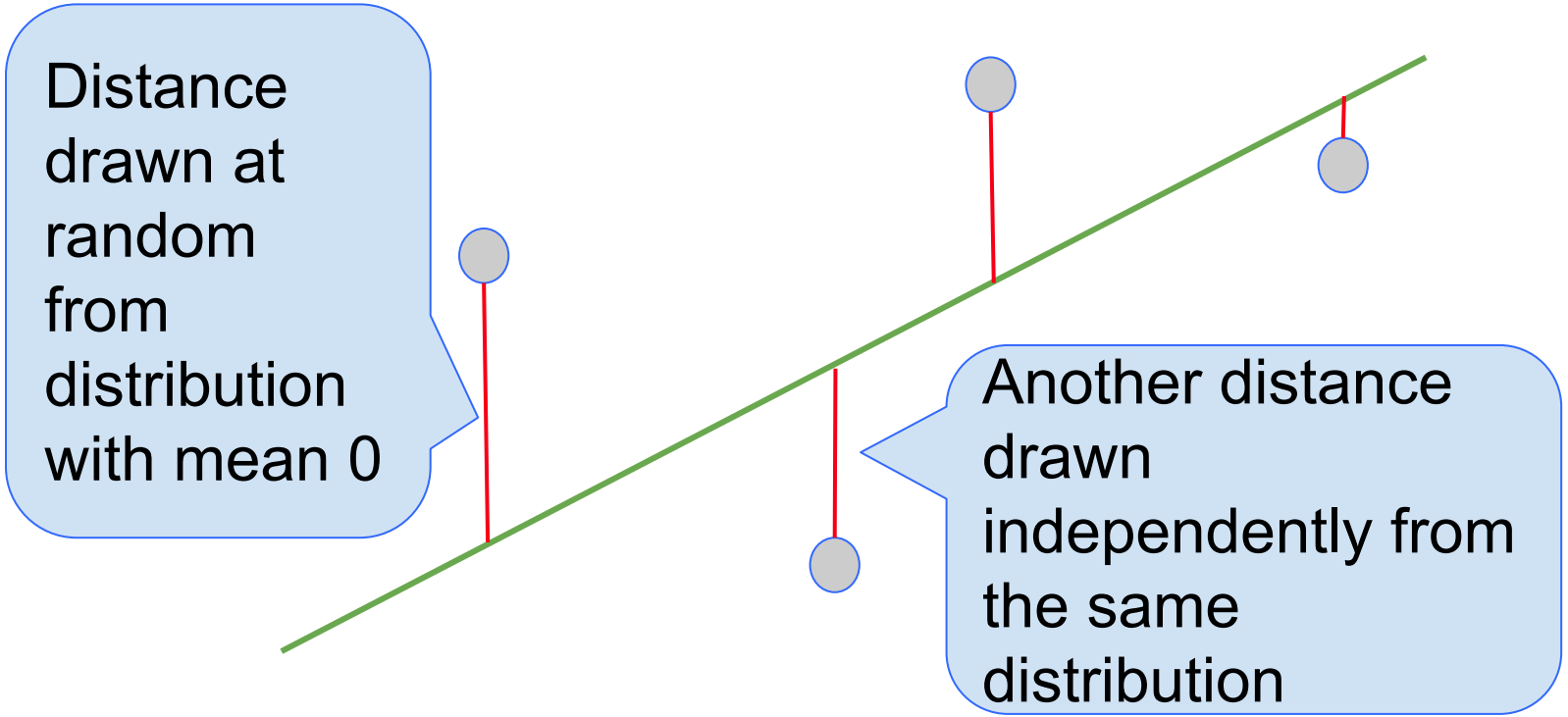
Regression Wrapup

Regression roadmap

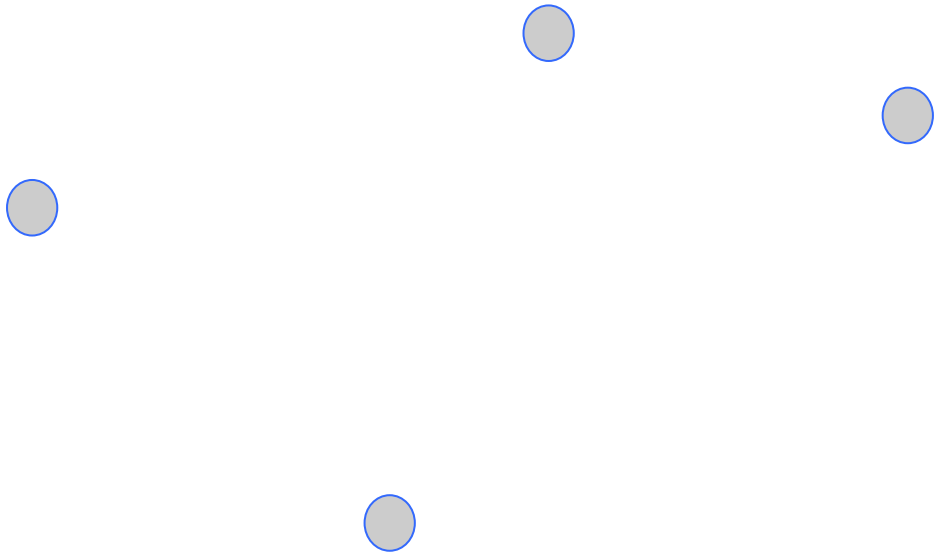
- Last Monday:
  - Association and correlation
- Last Wednesday
  - Prediction, scatterplots and lines
- Last Friday:
  - Least squares: finding the “best” line for a dataset
- Monday and Wednesday:
  - Residuals: analyzing mistakes and errors
- **Today**
  - **Regression inference and misc. topics**

Regression Model

A “Model”: Signal + Noise



What We Get to See



(Demo)

Prediction Variability

Regression Prediction

- **If the data come from the regression model,**
- **and if the sample is large, then:**

- The regression line is close to the true line
- Given a new value of *x*, predict *y* by finding the point on the regression line at that *x*

(Demo)

Confidence Interval for Prediction

- **Bootstrap the scatter plot**
- **Get a prediction for *y* using the regression line that goes through the resampled plot**
- Repeat the two steps above many times
- Draw the empirical histogram of all the predictions.
- Get the “middle 95%” interval.
- That’s an approximate 95% confidence interval for the height of the true line at *y*.

(Demo)

Predictions at Different Values of *x*

- Since *y* is correlated with *x*, the predicted values of *y* depend on the value of *x* (otherwise, there would be no point making predictions!)

- The width of the prediction's CI also depends on *x*.
  - Typically, when *x* is further away from its mean, the intervals for *y* is wider

(Demo)

The True Slope

Confidence Interval for True Slope

- **Bootstrap the scatter plot.**
- **Find the slope of the regression line through the bootstrapped plot.**
- Repeat.
- Draw the empirical histogram of all the generated slopes.
- Get the “middle 95%” interval.
- That’s an approximate 95% confidence interval for the slope of the true line.

(Demo)

Rain on the Regression Parade

We observed a slope based on our sample of points.

But what if the sample scatter plot got its slope just by chance?

What if the true line is actually FLAT?

Test Whether There Really is a Slope

- **Null hypothesis:** The slope of the true line is 0.
- **Alternative hypothesis:** No, it's not.
- Method:
  - Construct a bootstrap confidence interval for the true slope.
  - If the interval doesn't contain 0, the data are more consistent with the alternative
  - If the interval does contain 0, the data are more consistent with the null

(Demo)

Advanced Regression

Advanced Regression

- minimize() works no matter what*!

- Define a function that computes the prediction you want, then the error you want, for example:
  - Nonlinear functions of *x*
  - Multiple columns of the table for *x*
  - Other kinds of error instead of RMSE
- Nonlinear functions can get complicated, fast!

(Demo)

Prediction

Guessing the Value of an Attribute

- Based on incomplete information
- One way of making predictions:
  - To predict an outcome for an individual,
  - find others who are like that individual
  - and whose outcomes you know.
  - Use those outcomes as the basis of your prediction.

- Two Types of Prediction
  - Classification = Categorical; Regression = Numeric

Prediction Example: Spam or Not?

You made a Wells Fargo payment - wellsfargo.com You recently submitted a payment The ...

BUSINESS TRUST - -- I have a legal business proposal for you worth $23,000,000. If you kn...

Hi - Today???!!!! What a wonderful day! Congrats again! I am definitely not doing s...

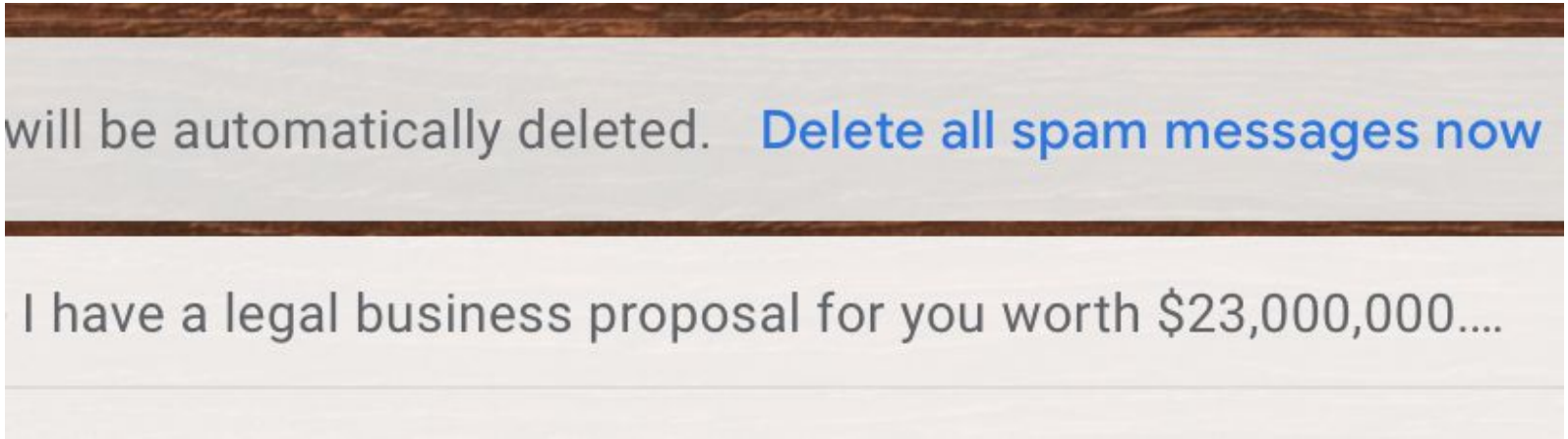
Michael Kors Handbags Up To 84% Plus Free Shipping! - Shop Handbags Online & In Store...

Machine Learning Algorithm

- A mathematical model
- calculated based on sample data ("training data")
- that makes predictions or decisions without being explicitly programmed to perform the task

Classification

Classification Examples



Classification Examples