Announcements

Regular lecture on Wed this week

Will finish grading in a couple of days

Group project “model” specs due by the 13th!
Easy and Hard Predictions
Why Accuracy and Error Are Misleading
Recap: Simulation v.s. Model Fitting

**Fitting**: Observe \( x_i, y_i \), infer \( \theta \)

Regression Problems \( \rightarrow \) Continuous \( Y \)

Classification Problems \( \rightarrow \) Discrete \( Y \)
Recap: Bias-Variance Tradeoff
Fraud Detection

**Task:** Build a model to determine whether a transaction at a bank is fraudulent or not.

Classification Problems -> Discrete Y

**Suppose:** 99.97% of transactions are honest and 0.03% are fraudulent.
How good is random(ish) guessing?

Look at some data-independent baselines

Suppose: 99.97% of transactions are honest and 0.03% are fraudulent.

Classifier: With 50% probability pick Fraud, 50% Not Fraud
(Bad) Idea 1. Even Guess

**Suppose:** 99.97% of transactions are honest and 0.03% are fraudulent.

**Classifier:** With 50% probability pick Fraud, 50% Not Fraud

What is the accuracy?

Accuracy = 50% = 0.5*0.9997 + 0.5*0.0003
(Bad) Idea 2. Best Guess

**Classifier:** With $p$ probability pick Class 1, $(1-p)$ Class 2

Accuracy = $p \times (\text{class 1 } \%) + (1-p) \times (\text{class 2 } \%)$

Accuracy = class 2 % + $p \times (\text{class 1 } \% - \text{class 2 } \%)$

Accuracy = $c_2 + p \times (c_1-c_2)$

Maximized at: $p = 1$ if $c_1 > c_2$

$p = 0$ if $c_2 > c_1$
How good is random(ish) guessing?

**Suppose:** 99.97% of transactions are honest and 0.03% are fraudulent.

**Classifier:** Pick highest likelihood class

99.97% Accuracy

Why is this a misleading statement?
How good is random(ish) guessing?

**Suppose:** 99.97% of transactions are honest and 0.03% are fraudulent.

**Classifier:** Pick highest likelihood class

- **100% Accurate**: True = Fraud
- **0% Accurate**: True = Not Fraud
False Positives v.s False Negatives

**False Positives:** Predict fraud when not fraud

**False Negatives:** Predict not fraud when actually fraud

Can tradeoff false positives for false negatives!
Logistic Regression

<table>
<thead>
<tr>
<th>X</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>[redness, how oval]</td>
<td></td>
</tr>
<tr>
<td>[0.95, 0.25]</td>
<td>0.96 &gt; a ~ 1</td>
</tr>
<tr>
<td>[0.25, 0.85]</td>
<td>0.1 &gt; a ~ 0</td>
</tr>
<tr>
<td>[0.45, 0.05]</td>
<td>0.87 &gt; a ~ 1</td>
</tr>
<tr>
<td>[0.05, 0.75]</td>
<td>0.14 &gt; a ~ 1</td>
</tr>
</tbody>
</table>

Set threshold \( a \) for desired tradeoff point
All Classifiers Have This Property

redness

ovalness

Sick

[0.95, 0.25]

Not Sick

[0.05, 0.75]

[0.25, 0.85]

Sick

[0.45, 0.05]
**Precision**: Fraction of truly detected among all detected.

\[ \frac{tp}{tp+fp} \]

**Recall**: Fraction of the total amount of detected that were actually detected.

\[ \frac{tp}{tp+fn} \]
**Precision:** Fraction of truly detected among all detected.

\[
\frac{\#\text{detect\_fraud}}{(\#\text{detect\_fraud} + \#\text{wrong})}
\]

**Recall:** Fraction of the total amount of detected that were actually detected.

\[
\frac{\#\text{detect\_fraud}}{(\#\text{detect\_fraud} + \#\text{missed})}
\]
**Precision**: Fraction of truly detected among all detected.

**False positive rate**

**Recall**: Fraction of the total amount of detected that were actually detected.

**False negative rate**
Visualization

Classification Threshold

Output of Logistic Regression model

0.0 1.0

TN TN TN TN TN TN TN TN TN TN FN FN FN FN FP TP TP TP TP TP TP TP TP

- Actually not spam
- Actually spam

Higher Recall

Lower Recall

Higher Precision

Lower Precision
Can flip them too!

**False Positives:** Predict *not fraud* when not fraud

**False Negatives:** Predict *fraud* when actually fraud

Can tradeoff false positives for false negatives!
Dealing With Imbalances

**Suppose:** 99.97% of transactions are honest and 0.03% are fraudulent.

**Measure:** Per-Class Precision and Recall

A good classifier is roughly balanced in P/R per class
Dependency on the Threshold?

Classification Threshold

0.0   Output of Logistic Regression model   1.0

Higher Recall

Lower Precision

Lower Recall

Higher Precision

- Actually not spam
- Actually spam
Precision Recall Curves

For all possible classification thresholds plot the P/R Curve
Precision Recall Curves

For all possible classification thresholds plot the P/R Curve

Random guessing!!!
Precision Recall Curves

For all possible classification thresholds plot the P/R Curve
Precision Recall Curves

Area-Under-Curve (AUC)

0.5 = Random Guessing

> 0.5 Better than random
Regression Problems

Measure errors in regression problems
1-parameter regression

Data-independent baselines

$$\min_c \sum_{i=1}^{N} (y_i - c)^2$$
1-parameter regression

\[ f(c) = \sum_{i=1}^{N} (y_i - c)^2 \]

\[ f'(c) = \sum_{i=1}^{N} -2(y_i - c) \]

\[ 0 = \sum_{i=1}^{N} -2(y_i - c) \]

\[ 0 = -2 \sum_{i=1}^{N} y_i + 2Nc \]

\[ c = \frac{1}{N} \sum_{i=1}^{N} y_i \]

C is the sample mean!!!
1-parameter regression

Best constant prediction is the mean of $y$!

Sample mean of $y$
1-parameter regression

What about the “error” in estimation?

\[ \sum_{i=1}^{N} (y_i - \mu_y)^2 \]

Proportional to the variance of \( y \)!
Benchmarking Regression

Compare regression results to a mean guess
R2 Score

Calculate the actual error compared to the error in a mean guess

\[ R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \]

Score = 1 (perfect prediction)
Score = 0 (no better than the mean)
Score < 0 (worse than the mean)
Explained Variation

\[
1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}
\]

“Explained” variation

How much of the variation in the data is explained by the predictor.
Explained Variation

$$1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

Calculate R2 Scores for different subsets of features.

Allows us to score the importance of different features
How much more predictive is the model if you additionally consider feature x over y?

$$R^2( \text{[featurex, featurey]} ) - R^2( \text{[featurey]} )$$
Explained Variation

\[ 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \]

Genetics explains 75% of human height.

Controlling for socio-economic status, age explains 80% of all cases.

\[ \ldots \]