Review

(training) instances → train a model → model

- supervised: label known
- unsupervised: label unknown

regression: continuous label
classification: discrete label
Q: What are the models (and how to learn them)?

(training) instances

supervised: label known
unsupervised: label unknown

train a model

model

regression: continuous label
classification: discrete label
<table>
<thead>
<tr>
<th></th>
<th>Regression</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised</td>
<td>Part 1</td>
<td>Part 2</td>
</tr>
<tr>
<td>Unsupervised</td>
<td></td>
<td>Part 3</td>
</tr>
</tbody>
</table>
1. Regression Task: Student GPA Prediction

\[ x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ \vdots \\ \vdots \end{pmatrix} = \begin{pmatrix} \text{major} \\ \#\text{hours} \\ \#\text{course} \\ \#\text{peer rej} \\ \vdots \\ \vdots \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 2 \\ 1 \\ \vdots \\ \vdots \end{pmatrix} \]

\[ y = 3.2 \]

Linear Regression Model

\[ f(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_p x_p \]
Example: Student GPA Prediction Model

A Model with One Feature

\[ f(x) = \theta_0 + \theta_1 x_1 \]
Graphical Interpretation of $f(x)$.

A Model with One Feature

$$f(x) = \theta_0 + \theta_1 x_1$$
Feature Space, Instance

A Model with One Feature

\( f(x) = \theta_0 + \theta_1 x_1 \)
Model, Bias, Slope

A Model with One Feature

$$f(x) = \theta_0 + \theta_1 x_1$$

- $y = $GPA
- $x_1 = $#hours
- $\theta_0$ is bias
- $\theta_1$ is slope
Label, Predicted Label

A Model with One Feature

\[ f(x) = \theta_0 + \theta_1 x_1 \]
Prediction Error

A Model with One Feature

\[ f(x) = \theta_0 + \theta_1 x_1 \]
Extension: a model with two features is a hyper-plan.

\[ y = \text{GPA} \]

\[ x_1 = \#\text{hw} \]

\[ x_2 = \text{major} \]

A Model with Two Feature

\[ f(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 \]
Q: how to learn $f(x)$ from training instances?

A Model with One Feature

$f(x) = \theta_0 + \theta_1 x_1$
Least-Square minimizes error on all training instances.

1. Least-Square method

\[ f = \min_f \sum_{(x,y)} [f(x) - y]^2 \]
Q: to void overfitting, can we restrict the domain of θ?

Example Models

\[ f_1(x) = 0.5 + 0.4x_1 + 0.21x_2 + \ldots + 0.19x_p \]

\[ f_2(x) = 0.5 + 17x_1 + 31x_2 + \ldots + 0.4x_p \]

1. Least-Square method

\[ f = \min_f \sum_{(x,y)} [f(x) - y]^2 \]
Ridge regression is LS plus regularization on $\theta$.

1. Least-Square method

$$f = \min_f \sum_{(x,y)} [f(x) - y]^2$$

2. Ridge Regression method

$$f = \min_f \sum_{(x,y)} [f(x) - y]^2 + \lambda \sum_{i=1}^{\theta} (\theta_i)^2$$

$\lambda$ is called regularization coefficient. It is a hyper-parameter.

Example Models

$$f_1(x) = 0.5 + 0.4x_1 + 0.21x_2 + ... + 0.19x_p$$

$$f_2(x) = 0.5 + 17x_1 + 31x_2 + ... + 0.4x_p$$
Q: to void overfitting, can we also use fewer features?

Example Models

1. Least-Square method

\[ f = \min_f \sum_{(x,y)} [f(x) - y]^2 \]

2. Ridge Regression method

\[ f = \min_f \sum_{(x,y)} [f(x) - y]^2 + \lambda \sum_{i=1}^k (\theta_i)^2 \]

\[
\begin{align*}
  f_1(x) &= 0.5 + 0.4x_1 + 0.21x_2 + \ldots + 0.19x_4 \\
  f_2(x) &= 0.5 + 0.7x_1 + 0.39x_4
\end{align*}
\]
LASSO is LS+automatic feature selection via regularization.

Example Models

\[ f_1(x) = 0.5 + 0.4x_1 + 0.21x_2 + \ldots + 0.19x_4 \]

\[ f_2(x) = 0.5 + 0.7x_1 + 0.39x_4 \]

1. Least-Square method
\[ f = \min_f \sum_{(x,y)}[f(x) - y]^2 \]

2. Ridge Regression method
\[ f = \min_f \sum_{(x,y)}[f(x) - y]^2 + \lambda \sum_{i=1}^{\theta_i} (\theta_i)^2 \]

3. LASSO method
\[ f = \min_f \sum_{(x,y)}[f(x) - y]^2 + \lambda \sum_{i=1}^{\theta_i} |\theta_i| \]

This term is L1-regularization.
Q: which model is more likely to be learned?

Example Models

1. Least-Square
   \[ f = \min_f \sum_{(x,y)} (f(x) - y)^2 \]

2. Ridge Regression
   \[ f = \min_f \sum_{(x,y)} (f(x) - y)^2 + \lambda \sum_{i=1}^{\theta} \theta_i^2 \]

3. LASSO
   \[ f = \min_f \sum_{(x,y)} (f(x) - y)^2 + \lambda \sum_{i=1}^{\theta} |\theta_i| \]

\[ f_1(x) = 0.1 + 0.7x_1 + 0.5x_2 + 0.6x_3 \]

\[ f_2(x) = 0.2 + 0.5x_1 + 0.3x_3 \]

\[ f_3(x) = 0.9 + 0.2x_1 + 0.3x_2 + 0.1x_3 \]
Ridge and Lasso do not penalize bias term.
Q: which result is more likely to be generated?

Example Models

\[ f_1(x) = 0.1 + 0.7x_1 + 0.5x_2 + 0.8x_3 \] (LS)

\[ f_2(x) = 0.2 + 0.5x_1 + 0.3x_3 \] (Lasso)

\[ f_3(x) = 0.9 + 0.2x_1 + 0.3x_2 + 0.4x_3 \] (Ridge)
Q: which learning method would you prefer to apply?

Our customer wants to build a GPA prediction model using all data given below.

<table>
<thead>
<tr>
<th>Student Name</th>
<th>Student ID</th>
<th>Steal</th>
<th>Lie, Cheat, Sneak</th>
<th>Behavior Problem</th>
<th>Peer Rejection</th>
<th>Low Academic Achievement</th>
<th>Negative Attitude</th>
<th>Aggressive Behavior</th>
<th>Total SRSS</th>
<th>GPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angel, Julio</td>
<td>2310</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>3.1</td>
</tr>
<tr>
<td>Akins, J'Monte</td>
<td>2013</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4.1</td>
</tr>
<tr>
<td>Backer, Brent</td>
<td>2031</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>9</td>
<td>2.3</td>
</tr>
</tbody>
</table>

1. LS \( \min_f \sum_{(x,y)} [f(x) - y]^2 \)

2. Ridge \( \min_f \sum_{(x,y)} [f(x) - y]^2 + \lambda \sum_{i=1}^{i} (\theta_i)^2 \)

3. LASSO \( \min_f \sum_{(x,y)} [f(x) - y]^2 + \lambda \sum_{i=1}^{i} |\theta_i| \)
Q: which learning method would you prefer to apply?

The customer wants to know which 2-3 attributes are critical for predicting GPA.

<table>
<thead>
<tr>
<th>Student Name</th>
<th>Student ID</th>
<th>Steal</th>
<th>Lie, Cheat, Sneak</th>
<th>Behavior Problem</th>
<th>Peer Rejection</th>
<th>Low Academic Achievement</th>
<th>Negative Attitude</th>
<th>Aggressive Behavior</th>
<th>Total SRSS</th>
<th>GPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angel, Julio</td>
<td>2310</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>3.1</td>
</tr>
<tr>
<td>Akins, J'Monte</td>
<td>2013</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4.1</td>
</tr>
<tr>
<td>Backer, Brent</td>
<td>2031</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>9</td>
<td>2.3</td>
</tr>
<tr>
<td>Boxwell, Kylie</td>
<td>2001</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2.5</td>
</tr>
<tr>
<td>Cartwright, Ashley</td>
<td>2152</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>3.2</td>
</tr>
<tr>
<td>Cox, Lucille</td>
<td>2002</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3.9</td>
</tr>
<tr>
<td>Hankins, Erin</td>
<td>2017</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>3.7</td>
</tr>
<tr>
<td>Illio, Helen</td>
<td>2132</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2.9</td>
</tr>
<tr>
<td>Jackson, Ronald</td>
<td>2003</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>12</td>
<td>1.7</td>
</tr>
<tr>
<td>Kemp, Patrice</td>
<td>2009</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>3.3</td>
</tr>
<tr>
<td>Parker, Stephanie</td>
<td>2004</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>2.7</td>
</tr>
<tr>
<td>Reed, Kent</td>
<td>2010</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3.6</td>
</tr>
<tr>
<td>Sterling, Michael</td>
<td>2022</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>2.4</td>
</tr>
<tr>
<td>Supervised</td>
<td>Regression</td>
<td>Classification</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td>------------------</td>
<td>----------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>linear regression</td>
<td>Part 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LS, Ridge, Lasso</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unsupervised</td>
<td></td>
<td>Part 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Q: What are the basic classification models?

Models

- naive bayes
- logistic regression
- k-nearest neighbor
- neural network
- support vector machine
- decision tree
1. Naive Bayes (Rule: if $\text{Pr}(y=\text{ham}|x) > \text{Pr}(y=\text{spam}|x)$, $x$ is ham.)

$$\text{Pr}(y = \text{ham} | x)$$

$$= \text{Pr}(y = \text{ham} | x_1, x_2, ..., x_p)$$

$$\propto \text{Pr}(x_1, x_2, ..., x_p | y=\text{ham}) \times \text{Pr}(y=\text{ham})$$

$$= \text{Pr}(x_1|y=\text{ham}) \times \text{Pr}(x_2|y=\text{ham}) \times ... \times \text{Pr}(y=\text{ham})$$

- assume features are independent
- estimate $\text{Pr}(x|y)$ from data
- same for $\text{Pr}( y=\text{spam} | x_1, x_2, ..., x_p)$
Q: is the feature independence assumption strong?

\[
Pr(y = \text{ham} \mid x) \\
= Pr(y = \text{ham} \mid x_1, x_2, \ldots, x_p) \\
\propto Pr(x_1, x_2, \ldots, x_p \mid y=\text{ham}) \ast Pr(y=\text{ham}) \\
= Pr(x_1|y=\text{ham}) \ast Pr(x_2|y=\text{ham}) \ast \cdots \ast Pr(y=\text{ham}) \\
- \text{assume features are independent} \\
- \text{estimate } Pr(x|y) \text{ from data} \\
- \text{same for } Pr( y=\text{spam} \mid x_1, x_2, \ldots, x_p)
We want to build a model to predict student GPA (A, B, C) based on their profiles.

Q: would you apply Naive Bayes?

<table>
<thead>
<tr>
<th>Student Name</th>
<th>Student ID</th>
<th>Steal</th>
<th>Lie, Cheat, Sneak</th>
<th>Behavior Problem</th>
<th>Peer Rejection</th>
<th>Low Academic Achievement</th>
<th>Negative Attitude</th>
<th>Aggressive Behavior</th>
<th>Total SRSS</th>
<th>GPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angel, Julio</td>
<td>2310</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>3.1</td>
</tr>
<tr>
<td>Akins, J'Monte</td>
<td>2013</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4.1</td>
</tr>
<tr>
<td>Backer, Brent</td>
<td>2031</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>9</td>
<td>2.3</td>
</tr>
<tr>
<td>Boxwell, Kylie</td>
<td>2001</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2.5</td>
</tr>
<tr>
<td>Cartwright, Ashley</td>
<td>2152</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>3.2</td>
</tr>
<tr>
<td>Cox, Lucille</td>
<td>2002</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3.9</td>
</tr>
<tr>
<td>Hankins, Erin</td>
<td>2017</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>3.7</td>
</tr>
<tr>
<td>Illio, Helen</td>
<td>2132</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2.9</td>
</tr>
<tr>
<td>Jackson, Ronald</td>
<td>2003</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>12</td>
<td>1.7</td>
</tr>
<tr>
<td>Kemp, Patrice</td>
<td>2009</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3.3</td>
</tr>
<tr>
<td>Parker, Stephanie</td>
<td>2004</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>2.7</td>
</tr>
<tr>
<td>Reed, Kent</td>
<td>2010</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3.6</td>
</tr>
<tr>
<td>Sterling, Michael</td>
<td>2022</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>2.4</td>
</tr>
</tbody>
</table>
Q: would you apply Naive Bayes?

We want to build a model to predict student GPA (A, B, C) based on their essays.

```
How I Plan to Stay Drug Free

Everyone knows drugs and alcohol are bad for you, but to some they just don't care. They don't care about their life, their family, even their own health. They don't see the big picture of how they destroy everything they touch, everything they help create, everything they say they love.
```
2. Logistic Regression (same rule as NB)

\[
Pr( y = \text{ham} | x) = \frac{1}{1 + \exp(-x^T\beta)}
\]

\[
Pr( y = \text{spam} | x) = \frac{\exp(-x^T\beta)}{1 + \exp(-x^T\beta)}
\]

- use logistic function to construct probability
- \(\beta\) is unknown vector and learned by MLE
- typically for binary classification
Q: can we apply LR to classify more than two classes?

We want to build a model to predict student GPA into A, B or C.
Strategy 1: **One-versus-All**
Relabel data for each subproblem.

train a LR f1

relabel as 0

relabel as 1
Strategy 2: **One-versus-One**
Q: how many subproblems in total? (1v1 or 1vA)
Get Final Prediction by **Majority Voting**

test data
Majority Voting Process

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>f1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Majority Voting Process

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>f1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>f2</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Majority Voting Process

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>f1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>f2</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>f3</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Majority Voting Process

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>f1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>f2</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>f3</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Sum</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>
Q: how about 1-versus-All?

one-vs.-rest

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>f1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>f2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>f3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sum</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
3. K-Nearest Neighbor

$\text{instance } x$ 

$\text{predicted label } f(x)$

K is the number of neighbors. It is a hyper-parameter.

Figure is from https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn.
Q: which K value gives higher model complexity?

\[ f(x) \]

instance \( x \)

predicted label \( f(x) \)

ham

spam

K=3

K=7
Q: would you use a larger K or a smaller K?

Q: difference between kNN and other models?

**Naive Bayes**

\[
\begin{align*}
\Pr( y = \text{ham} | x_1, x_2, ..., x_p ) &= ... \\
\Pr( y = \text{spam} | x_1, x_2, ..., x_p ) &= ...
\end{align*}
\]

**Logistic Regression**

\[
\begin{align*}
\Pr( y = \text{ham} | x ) &= 1 / [1 + \exp(-x^T\beta)] \\
\Pr( y = \text{spam} | x ) &= \exp(-x^T\beta) / [1 + \exp(-x^T\beta)]
\end{align*}
\]
4. Neural Network

A neuron is a linear function wrapped in a (non-linear) activation function. Weights are learned by back-propagation.

Figure is from https://databricks.com/glossary/neural-network.
Figure is from https://github.com/trekhleb/machine-learning-octave/tree/master/neural-network.
Parameters
1. network weights

Hyper-Parameters
1. # hidden layers
2. # neurons in each layer
3. type of activation function
Deep Neural Network

Figure is from https://www.kdnuggets.com/2017/12/deep-learning-made-easy-deep-cognition.html.
Q: which model has higher complexity?
Dropout to reduce overfitting.
DNN is good at extracting information from unstructured data.
DNN is good at extracting information from unstructured data.

Figure is from https://skymind.ai/wiki/neural-network.
5. SVM

a model is a decision boundary

predicted as red

predicted as green
Decision boundary can make mistakes.
Q: which boundary do you prefer?

f2  f3  f1
5. Support Vector Machine

optimal boundary has the max margin.
Support vectors are those lying on the margin. The optimal boundary has the max margin.
Soft-Margin SVM: a more tolerable SVM
Q: why tolerate some misbehaved instances?
Tolerance controlled by hyper-parameter C. 

large C, less tolerance
Q: what happens if C is infinitely large?

large C, less tolerance
Q: would you choose a larger C or smaller C?
Q: would you choose a larger C or smaller C?
Q: would you choose a larger C or smaller C?
6. Decision Tree

each internal node thresholds a feature.

each leaf node classifies arriving \( x \).

testing instance \( x \)

https://ese.wustl.edu/ContentFiles/Research/UndergraduateResearch/CompletedProjects/WebPages/sp14/SongSteimle/WebPage/classifiers.html
Generate a tree by splitting nodes (until they are ‘pure’).
Split a leaf node by thresholding one feature.

Does an email contain ‘lottery’?
Keep splitting impure leaf nodes (until child nodes are ‘pure’).

Does it contain ‘million’?
Determine labels for leaf nodes by majority voting.

assign spam

assign spam

assign spam

assign ham
Q1: how to measure purity of a node?
Entropy of a random variable measures its uncertainty.

Let $X$ and $Y$ be the hourly weather variables in Laramie and Cheyenne, respectively.

$X$: Cloud, Cloud, Overcast, Cloud, Sun, Cloud, Snow, Cloud, Cloud, ...

$Y$: Cloud, Sun, Rain, Snow, Cloud, Overcast, Rain, Rain, Sun, ...

Q: which weather variable has higher entropy?
Which **node** has higher entropy?
Q2: which **feature** should we threshold at a node?
Which feature would you prefer to threshold?

- 'million'?
- 'lottery'?
Q3: what hyper-parameter controls complexity of a tree?
Size of a tree is its hyper-parameter.

Example Hyper-parameters
1. Max Tree Depth
2. Max Tree Breadth
Q4: how to reduce overfitting of a tree?

Example hyper-parameters
1. Max Tree Depth
2. Max Tree Breadth
Q4: how to reduce overfitting of a tree?

Example hyper-parameters
1. Max Tree Depth
2. Max Tree Breadth

Reduce Overfitting
1. restrict size (during training)
2. pruning (after training)
An Example of Pruning

assign spam

assign spam

assign ham

assign spam

assign ham
Another Example of Pruning
Q: why do we want to prune the subtree?
Q: pros and cons of decision tree?

**Logistic Regression**

\[
\Pr(y=\text{ham}|x) = \frac{1}{1 + \exp(-x^T\beta)}
\]

\[
\Pr(y=\text{spam}|x) = \frac{\exp(-x^T\beta)}{1 + \exp(-x^T\beta)}
\]

**Naive Bayes**

\[
\Pr(y=\text{ham}|x_1, x_2, \ldots, x_p) = \ldots
\]

\[
\Pr(y=\text{spam}|x_1, x_2, \ldots, x_p) = \ldots
\]
Graphical Interpretation of Decision Tree

Figure is from “Classification and novel class detection in concept-drifting data streams under time constraints”.
Other Applications of Decision Tree

Decision Tree for PlayTennis

- Outlook
  - Sunny
  - Overcast
    - Yes
  - Rain

- Humidity
  - High
    - No
  - Normal
    - Yes

- Wind
  - Strong
  - Weak
    - Yes
Other Applications of Decision Tree

```
Type
  /    
Car   SUV
     /    
  Doors Minivan
     /    
  2     4
     /    
  +     -
     /    
  +     -
  /    
  -     +

Tires
  /    
Blackwall Whitewall
     /    
  +     -
```