Today:
- The Big Picture
- Overfitting
- Review: probability

Readings:
- Decision trees, overfitting
  - Mitchell, Chapter 3

Probability review
- Bishop Ch. 1 thru 1.2.3
- Bishop, Ch. 2 thru 2.2
- Andrew Moore’s online tutorial

Function Approximation: Decision Tree Learning

Problem Setting:
- Set of possible instances $X$
  - each instance $x$ in $X$ is a feature vector
    $x = <x_1, x_2, ..., x_n>$
- Unknown target function $f : X \rightarrow Y$
  - $Y$ is discrete valued
- Set of function hypotheses $H = \{ h | h : X \rightarrow Y \}$
  - each hypothesis $h$ is a decision tree

Input:
- Training examples $\{(x^{(i)}, y^{(i)})\}$ of unknown target function $f$

Output:
- Hypothesis $h \in H$ that best approximates target function $f$
Function approximation as Search for the best hypothesis

- ID3 performs heuristic search through space of decision trees

Function Approximation: The Big Picture

$H = \{ h : x \rightarrow y \}$

$X = \langle x_1, \ldots, x_n \rangle$

$|x| = 2^n$

$|H| = 2^{|H|} = 2^{|x|}$

How many labeled examples are needed in order to determine which of the $2^{|x|}$ hypotheses is the correct one?

All $2^{|x|}$ instances in $X$ must be labeled.

There is no free lunch!

Inductive inference - generalizing beyond the training data is impossible unless we add more assumptions (e.g., priors, world)
Which Tree Should We Output?

- ID3 performs heuristic search through space of decision trees
- It stops at smallest acceptable tree. Why?

Occam’s razor: prefer the simplest hypothesis that fits the data

Why Prefer Short Hypotheses? (Occam’s Razor)

Arguments in favor:

Arguments opposed:
Why Prefer Short Hypotheses? (Occam’s Razor)

Argument in favor:
• Fewer short hypotheses than long ones
  → a short hypothesis that fits the data is less likely to be a statistical coincidence
  → highly probable that a sufficiently complex hypothesis will fit the data

Argument opposed:
• Also fewer hypotheses containing a prime number of nodes and attributes beginning with “Z”
• What’s so special about “short” hypotheses?

Overfitting in Decision Trees

Consider adding noisy training example #15:
\(\text{Sunny, Hot, Normal, Strong, PlayTennis = No}\)

What effect on earlier tree?
Overfitting

Consider error of hypothesis $h$ over
- training data: $\text{error}_{\text{train}}(h)$
- entire distribution $\mathcal{D}$ of data: $\text{error}_{\mathcal{D}}(h)$

Hypothesis $h \in H$ overfits training data if there is an alternative hypothesis $h' \in H$ such that

$$\text{error}_{\text{train}}(h) < \text{error}_{\text{train}}(h')$$

and

$$\text{error}_{\mathcal{D}}(h) > \text{error}_{\mathcal{D}}(h')$$

Overfitting in Decision Tree Learning
Avoiding Overfitting

How can we avoid overfitting?

• stop growing when data split not statistically significant
• grow full tree, then post-prune

Avoiding Overfitting

How can we avoid overfitting?

• stop growing when data split not statistically significant
• grow full tree, then post-prune

How to select “best” tree:

• Measure performance over training data
• Measure performance over separate validation data set
• MDL: minimize 
  \[ size(tree) + size(misclassifications(tree)) \]
**Reduced-Error Pruning**

Split data into *training* and *validation* set

Create tree that classifies *training* set correctly

Do until further pruning is harmful:

1. Evaluate impact on *validation* set of pruning each possible node (plus those below it)
2. Greedily remove the one that most improves *validation* set accuracy

- produces smallest version of most accurate subtree
- What if data is limited?

---

**Effect of Reduced-Error Pruning**

![Graph showing the effect of reduced-error pruning on accuracy as a function of tree size.](image)
Rule Post-Pruning

1. Convert tree to equivalent set of rules
2. Prune each rule independently of others
3. Sort final rules into desired sequence for use

Perhaps most frequently used method (e.g., C4.5)

Converting A Tree to Rules
What you should know:

- Well posed function approximation problems:
  - Instance space, $X$
  - Sample of labeled training data $\{ \langle x^{(i)}, y^{(i)} \rangle \}$
  - Hypothesis space, $H = \{ f: X \to Y \}$

- Learning is a search/optimization problem over $H$
  - Various objective functions
    - minimize training error (0-1 loss)
    - among hypotheses that minimize training error, select smallest (?)
  - But inductive learning without some bias is futile!

- Decision tree learning
  - Greedy top-down learning of decision trees (ID3, C4.5, ...)
  - Overfitting and tree/rule post-pruning
  - Extensions…

Extra slides

extensions to decision tree learning
Continuous Valued Attributes

Create a discrete attribute to test continuous

- \( Temperature = 82.5 \)
- \( (Temperature > 72.3) = t, f \)

<table>
<thead>
<tr>
<th>Temperature</th>
<th>40</th>
<th>48</th>
<th>60</th>
<th>72</th>
<th>80</th>
<th>90</th>
</tr>
</thead>
<tbody>
<tr>
<td>PlayTennis</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

Attributes with Many Values

Problem:

- If attribute has many values, \( Gain \) will select it
- Imagine using \( Date = Jun.3.1996 \) as attribute

One approach: use \( GainRatio \) instead

\[
GainRatio(S, A) \equiv \frac{Gain(S, A)}{SplitInformation(S, A)}
\]

\[
SplitInformation(S, A) \equiv - \sum_{i=1}^{c} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}
\]

where \( S_i \) is subset of \( S \) for which \( A \) has value \( v_i \)
**Unknown Attribute Values**

What if some examples missing values of $A$?
Use training example anyway, sort through tree

- If node $n$ tests $A$, assign most common value of $A$ among other examples sorted to node $n$
- assign most common value of $A$ among other examples with same target value
- assign probability $p_i$ to each possible value $v_i$ of $A$
  - assign fraction $p_i$ of example to each descendant in tree

Classify new examples in same fashion

---

**Questions to think about (1)**

- ID3 and C4.5 are heuristic algorithms that search through the space of decision trees. Why not just do an exhaustive search?
Questions to think about (2)

• Consider target function $f: \langle x_1, x_2 \rangle \rightarrow y$, where $x_1$ and $x_2$ are real-valued, $y$ is boolean. What is the set of decision surfaces describable with decision trees that use each attribute at most once?

Questions to think about (3)

• Why use Information Gain to select attributes in decision trees? What other criteria seem reasonable, and what are the tradeoffs in making this choice?
The Problem of Induction

• David Hume (1711-1776): pointed out
  1. Empirically, induction seems to work
  2. Statement (1) is an application of induction.
• This stumped people for about 200 years
Probability Overview

- Events
  - discrete random variables, continuous random variables, compound events

- Axioms of probability
  - What defines a reasonable theory of uncertainty

- Independent events

- Conditional probabilities

- Bayes rule and beliefs

- Joint probability distribution

- Expectations

- Independence, Conditional independence

Random Variables

- Informally, A is a random variable if
  - A denotes something about which we are uncertain
  - perhaps the outcome of a randomized experiment

- Examples
  - A = True if a randomly drawn person from our class is female
  - A = The hometown of a randomly drawn person from our class
  - A = True if two randomly drawn persons from our class have same birthday

- Define P(A) as “the fraction of possible worlds in which A is true” or “the fraction of times A holds, in repeated runs of the random experiment”
  - the set of possible worlds is called the sample space, S
  - A random variable A is a function defined over S
    \[ A: S \rightarrow \{0,1\} \]
A little formalism

More formally, we have

• a **sample space** $S$ (e.g., set of students in our class)
  – aka the set of possible worlds

• a **random variable** is a function defined over the sample space
  – Gender: $S \rightarrow \{ m, f \}$
  – Height: $S \rightarrow$ Reals

• an **event** is a subset of $S$
  – e.g., the subset of $S$ for which Gender=f
  – e.g., the subset of $S$ for which (Gender=m) AND (eyeColor=blue)

• we’re often interested in probabilities of specific events
  • and of specific events conditioned on other specific events

Visualizing $A$

Sample space of all possible worlds

Its area is

$P(A) = \text{Area of reddish oval}$
The Axioms of Probability

• $0 \leq P(A) \leq 1$
• $P(\text{True}) = 1$
• $P(\text{False}) = 0$
• $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

[di Finetti 1931]:

when gambling based on “uncertainty formalism A” you can be exploited by an opponent

iff

your uncertainty formalism A violates these axioms
Interpreting the axioms

- $0 \leq P(A) \leq 1$
- $P(\text{True}) = 1$
- $P(\text{False}) = 0$
- $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

The area of $A$ can’t get any smaller than 0.
And a zero area would mean no world could ever have $A$ true.

The area of $A$ can’t get any bigger than 1.
And an area of 1 would mean all worlds will have $A$ true.
Interpreting the axioms

- $0 \leq P(A) \leq 1$
- $P(\text{True}) = 1$
- $P(\text{False}) = 0$
- $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

Theorems from the Axioms

- $0 \leq P(A) \leq 1$, $P(\text{True}) = 1$, $P(\text{False}) = 0$
- $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$
  \[ \Rightarrow P(\text{not } A) = P(\sim A) = 1 - P(A) \]
Theorems from the Axioms

- $0 \leq P(A) \leq 1$, $P(\text{True}) = 1$, $P(\text{False}) = 0$
- $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$
  $\Rightarrow P(\text{not } A) = P(\sim A) = 1 - P(A)$

$P(A \text{ or } \sim A) = 1$  $P(A \text{ and } \sim A) = 0$

$P(A \text{ or } \sim A) = P(A) + P(\sim A) - P(A \text{ and } \sim A)$

$1 = P(A) + P(\sim A) - 0$

Elementary Probability in Pictures

- $P(\sim A) + P(A) = 1$
Another useful theorem

- $0 \leq P(A) \leq 1$, $P(\text{True}) = 1$, $P(\text{False}) = 0$,
- $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

$\Rightarrow P(A) = P(A \land B) + P(A \land \neg B)$

$A = A \text{ and } (B \text{ or } \neg B) = (A \text{ and } B) \text{ or } (A \text{ and } \neg B)$

$P(A) = P(A \text{ and } B) + P(A \text{ and } \neg B) - P((A \text{ and } B) \text{ and } (A \text{ and } \neg B))$

$P(A) = P(A \text{ and } B) + P(A \text{ and } \neg B) - P(A \text{ and } A \text{ and } B \text{ and } \neg B)$

Elementary Probability in Pictures

- $P(A) = P(A \land B) + P(A \land \neg B)$
Multivalued Discrete Random Variables

- Suppose $A$ can take on more than 2 values
- $A$ is a random variable with arity $k$ if it can take on exactly one value out of $\{v_1, v_2, \ldots, v_k\}$
- Thus...
  \[ P(A = v_i \land A = v_j) = 0 \text{ if } i \neq j \]
  \[ P(A = v_1 \lor A = v_2 \lor \ldots \lor A = v_k) = 1 \]

Elementary Probability in Pictures

\[ \sum_{j=1}^{k} P(A = v_j) = 1 \]
Definition of Conditional Probability

\[
P(A \cap B) 
\]

\[
P(A|B) = \frac{P(A \cap B)}{P(B)} 
\]

Corollary: The Chain Rule

\[
P(A \cap B) = P(A|B) \cdot P(B) 
\]

Conditional Probability in Pictures

picture: \( P(B|A=2) \)
Independent Events

• Definition: two events $A$ and $B$ are independent if $\Pr(A \text{ and } B) = \Pr(A) \times \Pr(B)$

• Intuition: knowing $A$ tells us nothing about the value of $B$ (and vice versa)

Picture “$A$ independent of $B$”
Elementary Probability in Pictures
• let’s write 2 expressions for $P(A \cap B)$

\[ P(A \cap B) = P(B|A) \times P(A) \]


…by no means merely a curious speculation in the doctrine of chances, but necessary to be solved in order to a sure foundation for all our reasonings concerning past facts, and what is likely to be hereafter…. necessary to be considered by any that would give a clear account of the strength of analogical or inductive reasoning….


we call $P(A)$ the “prior”

and $P(A|B)$ the “posterior”
Other Forms of Bayes Rule

\[ P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\sim A)P(\sim A)} \]

\[ P(A|B \land X) = \frac{P(B|A \land X)P(A \land X)}{P(B \land X)} \]

You should know

- Events
  - discrete random variables, continuous random variables, compound events
- Axioms of probability
  - What defines a reasonable theory of uncertainty
- Independent events
- Conditional probabilities
- Bayes rule and beliefs
what does all this have to do with function approximation?

Your first consulting job

- A billionaire from the suburbs of Seattle asks you a question:
  - He says: I have thumbtack, if I flip it, what's the probability it will fall with the nail up?
  - You say: Please flip it a few times:

    ▼ ▼ ▼ ▼

  - You say: The probability is:
  - **He says**: Why???
  - You say: Because...
Thumbtack – Binomial Distribution

- \( P(\text{Heads}) = \theta, \ P(\text{Tails}) = 1-\theta \)

- \( \mathbf{D}: \downarrow \ \downarrow \ \downarrow \ \downarrow \ \downarrow \ \downarrow \)

- \( \downarrow \_1 \ \downarrow \_2 \ \downarrow \_3 \ \downarrow \_4 \ \downarrow \_5 \)

- Flips are i.i.d.:
  - Independent events
  - Identically distributed according to Binomial distribution

- Sequence \( D \) of \( \alpha_H \) Heads and \( \alpha_T \) Tails
  \[
P(D \mid \theta) = \theta^{\alpha_H} (1 - \theta)^{\alpha_T}
\]

Maximum Likelihood Estimation

- **Data:** Observed set \( D \) of \( \alpha_H \) Heads and \( \alpha_T \) Tails

- **Hypothesis:** Binomial distribution

- Learning \( \theta \) is an optimization problem
  - What’s the objective function?

- **MLE:** Choose \( \theta \) that maximizes the probability of observed data:
  \[
  \hat{\theta} = \arg \max_\theta \ P(D \mid \theta)
  = \arg \max_\theta \ \ln P(D \mid \theta)
  \]
Maximum Likelihood Estimate for $\Theta$

\[ \hat{\theta} = \arg \max_{\theta} \ln P(\mathcal{D} | \theta) \]

\[ = \arg \max_{\theta} \ln \theta^{\alpha_H} (1 - \theta)^{\alpha_T} \]

- Set derivative to zero:

\[ \frac{d}{d\theta} \ln P(\mathcal{D} | \theta) = 0 \]
How many flips do I need?

\[ \hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T} \]

Bayesian Learning

- Use Bayes rule:
  \[ P(\theta \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \theta)P(\theta)}{P(\mathcal{D})} \]

- Or equivalently:
  \[ P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta)P(\theta) \]
Beta prior distribution – P(θ)

\[ P(\theta) = \frac{\theta^{\beta_H - 1} (1 - \theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \sim \text{Beta}(\beta_H, \beta_T) \]

- Likelihood function: \( P(\mathcal{D} \mid \theta) = \theta^\alpha (1 - \theta)^{\alpha_T} \)
- Posterior: \( P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta) P(\theta) \)

Posterior distribution

- Prior: \( \text{Beta}(\beta_H, \beta_T) \)
- Data: \( \alpha_H \) heads and \( \alpha_T \) tails
- Posterior distribution:
\[ P(\theta \mid \mathcal{D}) \sim \text{Beta}(\beta_H + \alpha_H, \beta_T + \alpha_T) \]
Dirichlet distribution

- number of heads in N flips of a two-sided coin
  - follows a binomial distribution
  - Beta is a good prior (conjugate prior for binomial)

- what it’s not two-sided, but k-sided?
  - follows a multinomial distribution
  - Dirichlet distribution is the conjugate prior

\[
P(\theta_1, \theta_2, \ldots, \theta_K) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} \theta_i^{(\alpha_i - 1)}
\]
Estimating Parameters

- Maximum Likelihood Estimate (MLE): choose $\theta$ that maximizes probability of observed data $\mathcal{D}$
  \[
  \hat{\theta} = \arg \max_{\theta} P(\mathcal{D} | \theta)
  \]

- Maximum a Posteriori (MAP) estimate: choose $\theta$ that is most probable given prior probability and the data
  \[
  \hat{\theta} = \arg \max_{\theta} P(\theta | \mathcal{D})
  = \arg \max_{\theta} \frac{P(\mathcal{D} | \theta)P(\theta)}{P(\mathcal{D})}
  \]

You should know

- Probability basics
  - random variables, events, sample space, conditional probs, …
  - independence of random variables
  - Bayes rule
  - Joint probability distributions
  - calculating probabilities from the joint distribution

- Point estimation
  - maximum likelihood estimates
  - maximum a posteriori estimates
  - distributions – binomial, Beta, Dirichlet, …
Extra slides

The Joint Distribution

Recipe for making a joint distribution of M variables:

Example: Boolean variables A, B, C
The Joint Distribution

Recipe for making a joint distribution of $M$ variables:

1. Make a truth table listing all combinations of values of your variables (if there are $M$ Boolean variables then the table will have $2^M$ rows).

Example: Boolean variables $A, B, C$

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.30</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.05</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.10</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.05</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.05</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.10</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.25</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.10</td>
</tr>
</tbody>
</table>
The Joint Distribution

Recipe for making a joint distribution of M variables:

1. Make a truth table listing all combinations of values of your variables (if there are M Boolean variables then the table will have $2^M$ rows).
2. For each combination of values, say how probable it is.
3. If you subscribe to the axioms of probability, those numbers must sum to 1.

Example: Boolean variables A, B, C

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.30</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.05</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.10</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.05</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.05</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.10</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.25</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.10</td>
</tr>
</tbody>
</table>

Using the Joint

One you have the JD you can ask for the probability of any logical expression involving your attribute

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$
Using the Joint

\[ P(\text{Poor Male}) = 0.4654 \]

\[ P(E) = \sum_{\text{rows matching } E} P(\text{row}) \]

Using the Joint

\[ P(\text{Poor}) = 0.7604 \]

\[ P(E) = \sum_{\text{rows matching } E} P(\text{row}) \]
Inference with the Joint

\[
P(E_1 \mid E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)} = \frac{\sum P(\text{row})}{\sum P(\text{row})}
\]

\[P(\text{Male} \mid \text{Poor}) = \frac{0.4654}{0.7604} = 0.612\]
Expected values

Given discrete random variable $X$, the expected value of $X$, written $E[X]$ is

$$E[X] = \sum_{x \in \mathcal{X}} xP(X = x)$$

We also can talk about the expected value of functions of $X$

$$E[f(X)] = \sum_{x \in \mathcal{X}} f(x)P(X = x)$$

Covariance

Given two discrete r.v.’s $X$ and $Y$, we define the covariance of $X$ and $Y$ as

$$Cov(X, Y) = E[(X - E(X))(Y - E(Y))]$$

e.g., $X=$gender, $Y=$playsFootball
or $X=$gender, $Y=$leftHanded

Rememb

$$E[X] = \sum_{x \in \mathcal{X}} xP(X = x)$$
Example: Bernoulli model

- Data:
  - We observed $N$ iid coin tossing: $D = \{1, 0, ..., 0\}$

- Representation:
  
  Binary r.v.
  
  \[ x_i \in \{0, 1\} \]

- Model:
  
  \[
P(x) = \begin{cases} 
  1 - \theta & \text{for } x = 0 \\
  \theta & \text{for } x = 1 
\end{cases} \quad \Rightarrow \quad P(x) = \theta^x(1 - \theta)^{1-x}
\]

- How to write the likelihood of a single observation $x_i$?
  
  \[ P(x_i) = \theta^{x_i}(1 - \theta)^{1-x_i} \]

- The likelihood of dataset $D = \{x_1, ..., x_N\}$:
  
  \[
P(x_1, x_2, ..., x_N | \theta) = \prod_{i=1}^N P(x_i | \theta) = \prod_{i=1}^N \theta^{x_i}(1 - \theta)^{1-x_i} = \theta^{\sum x_i}(1 - \theta)^{N - \sum x_i} = \theta^{\text{success}}(1 - \theta)^{\text{failure}}
\]

You should know

- Probability basics
  - random variables, events, sample space, conditional probs, …
  - independence of random variables
  - Bayes rule
  - Joint probability distributions
  - calculating probabilities from the joint distribution

- Point estimation
  - maximum likelihood estimates
  - maximum a posteriori estimates
  - distributions – binomial, Beta, Dirichlet, …