Today:
• Probability
• Bayes Rule
• Estimating parameters
  • maximum likelihood
  • max a posteriori
many of these slides are derived from William Cohen, Andrew Moore, Aarti Singh, Eric Xing, Carlos Guestrin. - Thanks!

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

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: $S \rightarrow \text{Reals}$

• 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 1

P(A) = 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

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

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

\[
\begin{align*}
P(A \text{ or } \sim A) &= 1 \\
P(\sim A) + P(A) &= 1
\end{align*}
\]

Elementary Probability in Pictures

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

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

\[ \Rightarrow P(A) = P(A ^ B) + P(A ^ \sim B) \]

\[ A = [A \text{ and } (B \text{ or } \sim B)] = [(A \text{ and } B) \text{ or } (A \text{ and } \sim B)] \]

\[ P(A) = P(A \text{ and } B) + P(A \text{ and } \sim B) - P((A \text{ and } B) \text{ and } (A \text{ and } \sim B)) \]

\[ P(A) = P(A \text{ and } B) + P(A \text{ and } \sim B) - P(A \text{ and } B \text{ and } A \text{ and } \sim B) \]

Elementary Probability in Pictures

- \(P(A) = P(A ^ B) + P(A ^ \sim B)\)
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 \( \Pr(A \text{ and } B) = \Pr(A) \cdot \Pr(B) \)
- Intuition: knowing A tells us nothing about the value of B (and vice versa)

Visualizing Probabilities

- Sample space of all possible worlds
- Its area is 1
**Definition of Conditional Probability**

\[ P(A \cap B) \]

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

**Corollary: The Chain Rule**

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

\[ P(C \cap A \cap B) = P(C|A \cap B) P(A|B) P(B) \]
Independent Events

• Definition: two events A and B are *independent* if \( P(A \cap B) = P(A) \times P(B) \)

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

Bayes Rule

• let’s write 2 expressions for \( P(A \cap B) \)
\[ P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \]

Bayes’ rule

we call \( P(A) \) the “prior”

and \( P(A|B) \) the “posterior”


...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...

Other Forms of Bayes Rule

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

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

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

A = you have the flu,   B = you just coughed

Assume:
P(A) = 0.05
P(B\mid A) = 0.80
P(B\mid \sim A) = 0.2

what is P(flu \mid cough) = P(A\mid B)?

what does all this have to do with function approximation?
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></th>
<th></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>

[A. Moore]
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.

---

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>

---

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>

---

[A. Moore]
Using the Joint

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

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

[A. Moore]

Using the Joint

\[ P(\text{Poor Male}) = 0.4654 \quad P(E) = \sum_{\text{rows matching } E} P(\text{row}) \]

[A. Moore]
Using the Joint

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

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

Inference with the Joint

\[ P(E_1 \mid E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)} = \frac{\sum_{\text{rows matching } E_1 \text{ and } E_2} P(\text{row})}{\sum_{\text{rows matching } E_2} P(\text{row})} \]

\[ P(\text{Male} \mid \text{Poor}) = \frac{0.4654}{0.7604} = 0.612 \]
Learning and the Joint Distribution

Suppose we want to learn the function $f: \langle G, H \rangle \rightarrow W$

Equivalently, $P(W | G, H)$

Solution: learn joint distribution from data, calculate $P(W | G, H)$

e.g., $P(W=\text{rich} | G = \text{female}, H = 40.5^-) =$

[A. Moore]

---

<table>
<thead>
<tr>
<th>gender</th>
<th>hours_worked</th>
<th>wealth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>v0:40.5-</td>
<td>poor</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rich</td>
</tr>
<tr>
<td></td>
<td>v1:40.5+</td>
<td>poor</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rich</td>
</tr>
<tr>
<td>Male</td>
<td>v0:40.5-</td>
<td>poor</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rich</td>
</tr>
<tr>
<td></td>
<td>v1:40.5+</td>
<td>poor</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rich</td>
</tr>
</tbody>
</table>

sounds like the solution to learning $F: X \rightarrow Y$, or $P(Y | X)$.

Are we done?
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:
    \[ \uparrow \downarrow \uparrow \downarrow \]
    \[ \uparrow \downarrow \uparrow \downarrow \downarrow \]
  - You say: The probability is:
  - **He says: Why???
  - You say: Because…**

---

Thumbtack – Binomial Distribution

- \( P(\text{Heads}) = \theta, \ P(\text{Tails}) = 1-\theta \)
  
  D:
  \[ \uparrow \downarrow \uparrow \downarrow \downarrow \]
  \[ x_1, x_2, x_3, x_4, x_5 \]

  Flips produce data set \( D \) with \( \alpha_H \) heads and \( \alpha_T \) tails
  - Flips are independent, identically distributed 1’s and 0’s (Bernoulli)
  - \( \alpha_H \) and \( \alpha_T \) are counts that sum these outcomes (Binomial)

  \[ P(D|\theta) = P(\alpha_H, \alpha_T|\theta) = \theta^{\alpha_H} (1-\theta)^{\alpha_T} \]
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 $\Theta$

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

- Set derivative to zero:
  \[
  \frac{d}{d\theta} \ln P(D \mid \theta) = 0
  \]
Set derivative to zero: \[ \frac{d}{d\theta} \ln P(D | \theta) = 0 \]

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

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

---

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 – \( P(\theta) \)

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

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

- Likelihood function: $P(D | \theta) = \theta^\alpha H (1 - \theta)^\alpha T$
- Posterior: $P(\theta | D) \propto P(D | \theta) P(\theta)$

---

Posterior distribution

- Prior: $\text{Beta}(\beta_H, \beta_T)$
- Data: $\alpha_H$ heads and $\alpha_T$ tails
- Posterior distribution:
  $$P(\theta | D) \sim \text{Beta}(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

[C. Guestrin]
MAP for Beta distribution

\[ P(\theta \mid \mathcal{D}) = \frac{\theta^{\beta_H + \alpha_H - 1}(1 - \theta)^{\beta_T + \alpha_T - 1}}{B(\beta_H + \alpha_H, \beta_T + \alpha_T)} \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T) \]

- MAP: use most likely parameter:
  \[ \hat{\theta} = \arg \max_{\theta} P(\theta \mid \mathcal{D}) = \]

- Beta prior equivalent to extra thumbscrew flips
- As \( N \to \infty \), prior is “forgotten”
- But, for small sample size, prior is important!

Conjugate priors

- \( P(\theta) \) and \( P(\theta \mid D) \) have the same form

Eg. 1 Coin flip problem

Likelihood is \( \sim \) Binomial

\[ P(D \mid \theta) = \theta^{\alpha_H}(1 - \theta)^{\alpha_T} \]

If prior is Beta distribution,

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

Then posterior is Beta distribution

\[ P(\theta \mid D) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T) \]

For Binomial, conjugate prior is Beta distribution.

[C. Guestrin] [A. Singh]
Conjugate priors

• $P(\theta)$ and $P(\theta \mid D)$ have the same form

Eg. 2 Dice roll problem (6 outcomes instead of 2)

Likelihood is $\sim$ Multinomial($\theta = \{\theta_1, \theta_2, \ldots, \theta_k\}$)

$$P(D \mid \theta) = \theta_1^{\alpha_1} \theta_2^{\alpha_2} \ldots \theta_k^{\alpha_k}$$

If prior is Dirichlet distribution,

$$P(\theta) = \frac{\prod_{i=1}^{k} \theta_i^{\beta_i-1}}{B(\beta_1, \ldots, \beta_k)} \sim \text{Dirichlet}(\beta_1, \ldots, \beta_k)$$

Then posterior is Dirichlet distribution

$$P(\theta \mid D) \sim \text{Dirichlet}(\beta_1 + \alpha_1, \ldots, \beta_k + \alpha_k)$$

For Multinomial, conjugate prior is Dirichlet distribution.

---

Estimating Parameters

• Maximum Likelihood Estimate (MLE): choose $\hat{\theta}$ that maximizes probability of observed data $D$

$$\hat{\theta} = \arg \max_{\theta} \ P(D \mid \theta)$$

• Maximum a Posteriori (MAP) estimate: choose $\hat{\theta}$ that is most probable given prior probability and the data

$$\hat{\theta} = \arg \max_{\theta} \ P(\theta \mid D)$$

$$= \arg \max_{\theta} \ \frac{P(D \mid \theta)P(\theta)}{P(D)}$$
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

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

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

• Estimating parameters from data
  – maximum likelihood estimates (MLE)
  – maximum a posteriori estimates (MAP)
  – distributions – binomial, Beta, Dirichlet, …
  – conjugate priors
Expected values

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

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

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

Remember: \( E[X] = \sum_{x \in X} xP(X = x) \)

Example: Bernoulli model

- Data:
  - We observed \( \text{\textit{iid coin tossing}} \): \( D = \{1, 0, 1, \ldots, 0\} \)
- Representation:
  - Binary r.v.: \( x_\epsilon \{0,1\} \)
- 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 \( x_\epsilon \)?
    - \( P(x_\epsilon) = \theta^x(1 - \theta)^{1-x} \)
- The likelihood of dataset \( D = \{x_1, ..., x_N\} \):
  - \( P(x_1, x_2, ..., x_N \mid \theta) = \prod_{i=1}^{N} P(x_i \mid \theta) = \prod_{i=1}^{N} \theta^{x_i}(1 - \theta)^{1-x_i} = \theta^{\sum_{i=1}^{N} x_i} (1 - \theta)^{\sum_{i=1}^{N} (1 - x_i)} = \theta^{\text{success}} (1 - \theta)^{\text{failure}} \)