Probability Estimation

Machine Learning 10-601B
Seyoung Kim

Many of these slides are derived from Tom Mitchell, William Cohen, Eric Xing. Thanks!
Overview

• Joint probability distribution
  – A functional mapping $f: X \rightarrow Y$ via probability distribution

• Probability estimation
  – Maximum likelihood estimation
  – Maximum a priori estimation
What does all this have to do with function approximation for f: X->Y?
Joint Probability Distribution

Once you have the joint distribution, you can ask for the probability of any logical expression involving your attribute.
Using the Joint Distribution

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

\[
P(A) = P(A \land B) + P(A \land \sim B)
\]

\[
P(\text{Poor, Male}) = 0.4654
\]
Using the Joint Distribution

\[ P(Poor) = 0.7604 \]

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

<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>

\[
P(E_1 \mid E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)} = \frac{\sum P(\text{row})}{\sum P(\text{row})}\]
Inference with the Joint

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

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

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

Equivalently, \( P(W \mid G, H) \)

Solution: learn joint distribution from data, calculate \( P(W \mid G, H) \)

e.g., \( P(W=\text{rich} \mid G = \text{female}, H = 40.5- ) = \)
Density Estimation

- Our Joint Distribution learner is our first example of something called **Density Estimation**
- A Density Estimator learns a mapping from a set of attributes values to a Probability
Density Estimation

• Compare it against the two other major kinds of models:

  Input Attributes → Density Estimator → Probability

  Input Attributes → Classifier → Prediction of categorical output or class
  One of a few discrete values

  Input Attributes → Regressor → Prediction of real-valued output
Density Estimation ➔ Classification

To classify \( x \)
1. Use your estimator to compute \( \hat{P}(x, y_1), \ldots, \hat{P}(x, y_k) \)
2. Return the class \( y^* \) with the highest predicted probability

Ideally is correct with

\[
\hat{P}(y^* \mid x) = \frac{\hat{P}(x, y^*)}{\hat{P}(x, y_1) + \ldots + \hat{P}(x, y_k)}
\]

Binary case: predict POS if \( \hat{P}(x, y_{pos}) > 0.5 \)
Classification vs Density Estimation

Classification

Density Estimation
Classification vs density estimation
Modeling Uncertainty with Probabilities

• Y is a Boolean-valued random variable if
  – Y denotes an event,
  – there is uncertainty as to whether Y occurs.
• More examples
  – Y = You wake up tomorrow with a headache
  – Y = The US president in 2023 will be male
  – Y = there is intelligent life elsewhere in our galaxy
  – Y = the 1,000,000,000,000\textsuperscript{th} digit of π is 7
  – Y = I woke up today with a headache
• Define P(Y|X) as “the fraction of possible worlds in which Y is true, given X”
sounds like the solution to learning

\[ F: X \rightarrow Y, \]

or \[ P(Y \mid X). \]

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:

☐ You say: The probability is:

☐ **He says: Why???
☐ You say: Because…

[C. Guestrin]
Thumbtack – Binomial Distribution

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

\[ D = \{ D_1, D_2, D_3, D_4, D_5 \} \]

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)

\[
P(D|\theta) = P(\alpha_H, \alpha_T|\theta) = \theta^{\alpha_H} (1 - \theta)^{\alpha_T} \left( \frac{\alpha_H + \alpha_T}{\alpha_H} \right)
\]

[C. Guestrin]
Maximum Likelihood Estimation

- **Data:** Observed set $D$ of $\alpha_H$ Heads and $\alpha_T$ Tails
- **Hypothesis:** Binomial distribution $P(D|\theta) = P(\alpha_H, \alpha_T|\theta) = \theta^{\alpha_H} (1 - \theta)^{\alpha_T} \binom{\alpha_H + \alpha_T}{\alpha_H}$
- Learning $\theta$ is an optimization problem
  - What’s the objective function?

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

[C. Guestrin]
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:

$$\frac{d}{d\theta} \ln P(D \mid \theta) = 0$$

[C. Guestrin]
\hat{\theta} = \arg \max_{\theta} \ln P(D \mid \theta)

= \arg \max_{\theta} \ln \theta^\alpha H (1 - \theta)^\alpha T

\text{Set derivative to zero:}

\frac{d}{d\theta} \ln P(D \mid \theta) = 0

[C. Guestrin]
How many flips do I need?

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

I bought a loaded 20-faced die (d20) on EBay...but it didn’t come with any specs. How can I find out how it behaves?

1. Collect some data (20 rolls)
2. Estimate $P(i) = \frac{\text{CountOf(rolls of } i)}{\text{CountOf(any roll)}}$
Issues with MLE estimate

I bought a loaded d20 on EBay...but it didn’t come with any specs. How can I find out how it behaves?

But: Do I really think it’s *impossible* to roll a 1, 2 or 3?
A better solution

I bought a loaded d20 on EBay...but it didn’t come with any specs. How can I find out how it behaves?

0. Imagine some data (20 rolls, each i shows up 1x)
1. Collect some data (20 rolls)
2. Estimate P(i)
A better solution?

\[ P(i) = \frac{\text{CountOf}(i) + 1}{\text{CountOf}(\text{ANY}) + \text{CountOf}(\text{IMAGINED})} \]

\[
\begin{align*}
P(1) &= 1/40 \\
P(2) &= 1/40 \\
P(3) &= 1/40 \\
P(4) &= (2+1)/40 \\
\vdots \\
P(19) &= (5+1)/40 \\
P(20) &= (4+1)/40 = 1/8 
\end{align*}
\]

MAP = maximum a posteriori estimate

0.2 vs. 0.125 – really different! Maybe I should “imagine” less data?
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 – $P(\theta)$

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

- Posterior: $P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta)P(\theta)$
Posterior distribution

- Prior: $Beta(\beta_H, \beta_T)$
- Data: $\alpha_H$ heads and $\alpha_T$ tails
- Posterior distribution:
  $$P(\theta \mid D) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$
MAP for Beta distribution

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

- MAP: use most likely parameter:
  \[ \hat{\theta} = \arg \max_{\theta} P(\theta \mid \mathcal{D}) = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H - 1 + \alpha_T + \beta_T - 1} \]

- Beta prior equivalent to extra thumbtack flips
- As \( N \to \infty \), prior is “forgotten”
- But, for small sample size, prior is important!  
  [C. Guestrin]
Conjugate priors

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

Eg. 1 Coin flip problem

Likelihood is $\sim$ Binomial

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

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

$$P(\theta|D) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$

For Binomial, conjugate prior is Beta distribution.

[A. Singh]
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

$$P(\theta_1, \theta_2, ... \theta_K) = \frac{1}{B(\alpha)} \prod_{i} \theta_i^{(\alpha_i-1)}$$
Conjugate priors

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

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

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

\[
P(D | \theta) = \theta_1^{\alpha_1} \theta_2^{\alpha_2} \ldots \theta_k^{\alpha_k}
\]

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

\[
P(\theta|D) \sim \text{Dirichlet}(\beta_1 + \alpha_1, \ldots, \beta_k + \alpha_k)
\]

For Multinomial, conjugate prior is Dirichlet distribution.

[A. Singh]
Estimating Parameters

- Maximum Likelihood Estimate (MLE): choose \( \theta \) that maximizes the probability of observed data

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

- Maximum a Posteriori (MAP) estimate: choose \( \theta \) that is most probable given prior probability and the data

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

Remember:
$$E[X] = \sum_{x \in \mathcal{X}} x P(X = x)$$
You should know

• Density estimation and its relation to classification

• Estimating parameters from data
  – maximum likelihood estimates
  – maximum a posteriori estimates
  – distributions – binomial, Beta, Dirichlet, ...
  – conjugate priors