MLE/MAP
Q: Professor Gormley said there might be an error in the corollaries of the Realizable / Agnostic case for infinite $|H|$. What are the correct versions?

A: Here they are...

**Corollary 3 (Realizable, Infinite $|H|$).** For some $\delta > 0$, with probability at least $(1 - \delta)$, for any hypothesis $h$ in $\mathcal{H}$ consistent with the data (i.e. with $\hat{R}(h) = 0$),

$$R(h) \leq O \left( \frac{1}{N} \left[ \text{VC}(|\mathcal{H}|) \ln \left( \frac{N}{\text{VC}(|\mathcal{H}|)} \right) + \ln \left( \frac{1}{\delta} \right) \right] \right) \quad (1)$$

**Corollary 4 (Agnostic, Infinite $|\mathcal{H}|$).** For some $\delta > 0$, with probability at least $(1 - \delta)$, for all hypotheses $h$ in $\mathcal{H}$,

$$R(h) \leq \hat{R}(h) + O \left( \sqrt{\frac{1}{N} \left[ \text{VC}(|\mathcal{H}|) + \ln \left( \frac{1}{\delta} \right) \right]} \right) \quad (2)$$
Reminders

• Homework 6: PAC Learning / Generative Models
  – Out: Mon, Mar 26 (+/-)
  – Due: Mon, Apr 02 (+/-) at 11:59pm
PROBABILITY
# Random Variables: Definitions

<table>
<thead>
<tr>
<th>Discrete Random Variable</th>
<th>$X$</th>
<th>Random variable whose values come from a countable set (e.g. the natural numbers or {True, False})</th>
</tr>
</thead>
</table>
| Probability mass function (pmf) | $p(x)$ | Function giving the probability that discrete r.v. $X$ takes value $x$. \[
    p(x) := P(X = x)
\] |
## Random Variables: Definitions

<table>
<thead>
<tr>
<th>Continuous Random Variable</th>
<th>$X$</th>
<th>Random variable whose values come from an interval or collection of intervals (e.g. the real numbers or the range $(3, 5)$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability density function (pdf)</td>
<td>$f(x)$</td>
<td>Function the returns a nonnegative real indicating the relative likelihood that a continuous r.v. $X$ takes value $x$</td>
</tr>
</tbody>
</table>

- For any continuous random variable: $P(X = x) = 0$
- Non-zero probabilities are only available to intervals:

$$P(a \leq X \leq b) = \int_{a}^{b} f(x) \, dx$$
Random Variables: Definitions

<table>
<thead>
<tr>
<th>Cumulative distribution function</th>
<th>$F(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Function that returns the probability that a random variable $X$ is less than or equal to $x$:</td>
<td>$F(x) = P(X \leq x)$</td>
</tr>
</tbody>
</table>

- For **discrete** random variables:

  $F(x) = P(X \leq x) = \sum_{x' < x} P(X = x') = \sum_{x' < x} p(x')$

- For **continuous** random variables:

  $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x')dx'$
Notational Shortcuts

A convenient shorthand:

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

\[ \Rightarrow \text{ For all values of } a \text{ and } b: \]

\[ P(A = a|B = b) = \frac{P(A = a, B = b)}{P(B = b)} \]
Notational Shortcuts

But then how do we tell $P(E)$ apart from $P(X)$?

Instead of writing:

$$P(A|B) = \frac{P(A, B)}{P(B)}$$

We should write:

$$P_{A|B}(A|B) = \frac{P_{A,B}(A, B)}{P_B(B)}$$

...but only probability theory textbooks go to such lengths.
COMMON PROBABILITY DISTRIBUTIONS
Common Probability Distributions

• For Discrete Random Variables:
  – Bernoulli
  – Binomial
  – Multinomial
  – Categorical
  – Poisson

• For Continuous Random Variables:
  – Exponential
  – Gamma
  – Beta
  – Dirichlet
  – Laplace
  – Gaussian (1D)
  – Multivariate Gaussian
Common Probability Distributions

Beta Distribution

Probability density function:

\[ f(\phi|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1 - x)^{\beta-1} \]
Common Probability Distributions

Dirichlet Distribution

probability density function:

\[
f(\phi|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1 - x)^{\beta-1}
\]
Common Probability Distributions

Dirichlet Distribution

probability density function:

\[ p(\vec{\phi} | \alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} \phi_k^{\alpha_k - 1} \]

where \( B(\alpha) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K} \alpha_k)} \)
EXPECTATION AND VARIANCE
The **expected value** of $X$ is $E[X]$. Also called the mean.

**Discrete random variables:**

Suppose $X$ can take any value in the set $\mathcal{X}$.

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

**Continuous random variables:**

$$E[X] = \int_{-\infty}^{+\infty} xf(x) \, dx$$
The variance of $X$ is $Var(X)$.

$$Var(X) = E[(X - E[X])^2]$$

- **Discrete random variables:**
  $$Var(X) = \sum_{x \in \mathcal{X}} (x - \mu)^2 p(x)$$

- **Continuous random variables:**
  $$Var(X) = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x) dx$$
Joint probability
Marginal probability
Conditional probability

MULTIPLE RANDOM VARIABLES
Joint Probability

- Key concept: two or more random variables may interact. Thus, the probability of one taking on a certain value depends on which value(s) the others are taking.
- We call this a joint ensemble and write
  \[ p(x, y) = \text{prob}(X = x \text{ and } Y = y) \]
Marginal Probabilities

- We can "sum out" part of a joint distribution to get the marginal distribution of a subset of variables:

\[ p(x) = \sum_{y} p(x, y) \]

- This is like adding slices of the table together.

- Another equivalent definition: \( p(x) = \sum_{y} p(x|y)p(y) \).
Conditional Probability

- If we know that some event has occurred, it changes our belief about the probability of other events.
- This is like taking a ”slice” through the joint table.

\[ p(x|y) = \frac{p(x, y)}{p(y)} \]
Independence and Conditional Independence

- Two variables are independent iff their joint factors:

\[ p(x, y) = p(x)p(y) \]

- Two variables are conditionally independent given a third one if for all values of the conditioning variable, the resulting slice factors:

\[ p(x, y|z) = p(x|z)p(y|z) \quad \forall z \]
MLE AND MAP
Suppose we have data $\mathcal{D} = \{x^{(i)}\}_{i=1}^{N}$

**Principle of Maximum Likelihood Estimation:**
Choose the parameters that maximize the likelihood of the data.

$$\theta^{\text{MLE}} = \arg\max_{\theta} \prod_{i=1}^{N} p(x^{(i)}|\theta)$$

Maximum Likelihood Estimate (MLE)
MLE

What does maximizing likelihood accomplish?
• There is only a finite amount of probability mass (i.e. sum-to-one constraint)
• MLE tries to allocate as much probability mass as possible to the things we have observed...

... at the expense of the things we have not observed
MLE

Example: MLE of Exponential Distribution

- pdf of Exponential(λ): \( f(x) = \lambda e^{-\lambda x} \)
- Suppose \( X_i \sim \text{Exponential}(\lambda) \) for \( 1 \leq i \leq N \).
- Find MLE for data \( D = \{ x^{(i)} \}_{i=1}^{N} \)
- First write down log-likelihood of sample.
- Compute first derivative, set to zero, solve for \( \lambda \).
- Compute second derivative and check that it is concave down at \( \lambda^{MLE} \).
Example: MLE of Exponential Distribution

- First write down log-likelihood of sample.

\[ \ell(\lambda) = \sum_{i=1}^{N} \log f(x^{(i)}) \quad (1) \]

\[ = \sum_{i=1}^{N} \log(\lambda \exp(-\lambda x^{(i)})) \quad (2) \]

\[ = \sum_{i=1}^{N} \log(\lambda) + -\lambda x^{(i)} \quad (3) \]

\[ = N \log(\lambda) - \lambda \sum_{i=1}^{N} x^{(i)} \quad (4) \]
Example: MLE of Exponential Distribution

- Compute first derivative, set to zero, solve for $\lambda$.

\[
\frac{d\ell(\lambda)}{d\lambda} = \frac{d}{d\lambda} N \log(\lambda) - \lambda \sum_{i=1}^{N} x^{(i)} \tag{1}
\]

\[
= \frac{N}{\lambda} - \sum_{i=1}^{N} x^{(i)} = 0 \tag{2}
\]

\[
\Rightarrow \lambda^{\text{MLE}} = \frac{N}{\sum_{i=1}^{N} x^{(i)}} \tag{3}
\]
Example: MLE of Exponential Distribution

- pdf of Exponential(λ): \( f(x) = \lambda e^{-\lambda x} \)
- Suppose \( X_i \sim \text{Exponential}(\lambda) \) for \( 1 \leq i \leq N \).
- Find MLE for data \( D = \{x^{(i)}\}_{i=1}^N \)
- First write down log-likelihood of sample.
- Compute first derivative, set to zero, solve for \( \lambda \).
- Compute second derivative and check that it is concave down at \( \lambda^{\text{MLE}} \).
In-Class Exercise

Show that the MLE of parameter p for N samples drawn from Bernoulli(p) is:

\[ \hat{p}_{MLE} = \frac{\text{Number of } x_i = 1}{N} \]

Steps to answer:

1. Write log-likelihood of sample
2. Compute derivative w.r.t. p
3. Set derivative to zero and solve for p
Learning from Data (Frequentist)

Whiteboard

– Optimization for MLE
– Examples: 1D and 2D optimization
– Example: MLE of Bernoulli
– Example: MLE of Categorical
– Aside: Method of Lagrange Multipliers
MLE vs. MAP

Suppose we have data $\mathcal{D} = \{x^{(i)}\}_{i=1}^{N}$.

**Principle of Maximum Likelihood Estimation:**
Choose the parameters that maximize the likelihood of the data.

$$\theta^{\text{MLE}} = \arg\max_{\theta} \prod_{i=1}^{N} p(x^{(i)} | \theta)$$

Maximum Likelihood Estimate (MLE)

**Principle of Maximum a posteriori (MAP) Estimation:**
Choose the parameters that maximize the posterior of the parameters given the data.

$$\theta^{\text{MAP}} = \arg\max_{\theta} \prod_{i=1}^{N} p(\theta | x^{(i)})$$

Maximum a posteriori (MAP) estimate
MLE vs. MAP

Suppose we have data \( \mathcal{D} = \{x^{(i)}\}_{i=1}^{N} \)

**Principle of Maximum Likelihood Estimation:** Choose the parameters that maximize the likelihood of the data.

\[
\theta^{\text{MLE}} = \arg\max_{\theta} \prod_{i=1}^{N} p(x^{(i)}|\theta)
\]

Maximum Likelihood Estimate (MLE)

**Principle of Maximum a posteriori (MAP) Estimation:** Choose the parameters that maximize the posterior of the parameters given the data.

\[
\theta^{\text{MAP}} = \arg\max_{\theta} \prod_{i=1}^{N} p(x^{(i)}|\theta)p(\theta)
\]

Maximum a posteriori (MAP) estimate
Learning from Data (Bayesian)

Whiteboard

– *maximum a posteriori* (MAP) estimation
– Optimization for MAP
– Example: MAP of Bernoulli—Beta
Takeaways

• One view of what ML is trying to accomplish is **function approximation**

• The principle of **maximum likelihood estimation** provides an alternate view of learning

• **Synthetic data** can help **debug** ML algorithms

• Probability distributions can be used to **model** real data that occurs in the world (don’t worry we’ll make our distributions more interesting soon!)
Learning Objectives

**MLE / MAP**

*You should be able to...*

1. Recall probability basics, including but not limited to: discrete and continuous random variables, probability mass functions, probability density functions, events vs. random variables, expectation and variance, joint probability distributions, marginal probabilities, conditional probabilities, independence, conditional independence

2. Describe common probability distributions such as the Beta, Dirichlet, Multinomial, Categorical, Gaussian, Exponential, etc.

3. State the principle of maximum likelihood estimation and explain what it tries to accomplish

4. State the principle of maximum a posteriori estimation and explain why we use it

5. Derive the MLE or MAP parameters of a simple model in closed form