Lecture 7:
- Maximum Likelihood Estimation (MLE)
- Maximum a Posteriori (MAP)
Administrative

- **Assignment 2** will be out on Thursday
  - It is due **November 2** (i.e. in 2 weeks)
  - You will implement
    - Naive Bayes classifier for sentiment analysis
Administrative

- **Project proposal** due October 30
- A half page description
  - problem to be investigated,
  - why it is interesting,
  - what data you will use,
  - related work.
Today

• Probabilities
  - Dependence, Independence, Conditional Independence

• Parameter estimation
  - Maximum Likelihood Estimation (MLE)
  - Maximum a Posteriori (MAP)
Def: A sample space $\Omega$ is the set of all possible outcomes of a (conceptual or physical) random experiment. ($\Omega$ can be finite or infinite.)

Examples:
- $\Omega$ may be the set of all possible outcomes of a dice roll (1,2,3,4,5,6)
- Pages of a book opened randomly. (1-157)
- Real numbers for temperature, location, time, etc
Last time... Events

We will ask the question:
What is the probability of a particular event?

Def: Event $A$ is a **subset** of the sample space $\Omega$

Examples:
What is the probability of
- the book is open at an odd number
- rolling a dice the number $<4$
- a random person’s height $X : a<X<b$
Def: Probability $P(A)$, the probability that event (subset) $A$ happens, is a function that maps the event $A$ onto the interval $[0, 1]$. $P(A)$ is also called the probability measure of $A$.

Example: What is the probability that the number on the dice is 2 or 4?
(i) Nonnegativity: \( P(A) \geq 0 \) for each \( A \) event.

(ii) \( P(\Omega) = 1 \).

(iii) \( \sigma \)-additivity: For disjoint sets (events) \( A_i \), we have

\[
P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)
\]

Consequences:

\[
P(\emptyset) = 0.
\]

\[
P(A \cup B) = P(A) + P(B) - P(A \cap B).
\]

\[
P(A^c) = 1 - P(A).
\]
Last time... Venn Diagram

\[ P(A \cup B) = P(A) + P(B) - P(A \cap B) \]
**Def:** Real valued random variable is a function of the outcome of a randomized experiment

\[ X : \Omega \rightarrow \mathbb{R} \]

where

\[
P(a < X < b) \doteq P(\omega : a < X(\omega) < b)
\]

\[
P(X = a) \doteq P(\omega : X(\omega) = a)
\]

**Examples:**

- **Discrete random variable examples (\(\Omega\) is discrete):**
  - \(X(\omega) = \text{True if a randomly drawn person (\(\omega\)) from our class (\(\Omega\)) is female}\)
  - \(X(\omega) = \text{The hometown } X(\omega) \text{ of a randomly drawn person (\(\omega\)) from our class (\(\Omega\))}\)
Last time... Discrete Distributions

- Bernoulli distribution: $\text{Ber}(p)$

  $\Omega = \{\text{head, tail}\}$  $X(\text{head}) = 1$, $X(\text{tail}) = 0$. 
Last time... Discrete Distributions

- Bernoulli distribution: $\text{Ber}(p)$

\[ \Omega = \{\text{head, tail}\} \quad X(\text{head}) = 1, \quad X(\text{tail}) = 0. \]

\[ P(X = a) = P(\omega : X(\omega) = a) = \begin{cases} p, & \text{for } a = 1 \\ 1 - p, & \text{for } a = 0 \end{cases} \]
Last time… Discrete Distributions

• Bernoulli distribution: Ber(p)

$$\Omega = \{\text{head, tail}\} \quad X(\text{head}) = 1, \quad X(\text{tail}) = 0.$$ 

$$P(X = a) = P(\omega : X(\omega) = a) = \begin{cases} p, & \text{for } a = 1 \\ 1 - p, & \text{for } a = 0 \end{cases}$$

• Binomial distribution: Bin(n,p)

Suppose a coin with head prob. $p$ is tossed $n$ times. What is the probability of getting $k$ heads and $n-k$ tails?

$$\Omega = \{\text{possible } n \text{ long head/tail series}\}, \quad |\Omega| = 2^n$$

$$K(\omega) = \text{number of heads in } \omega = (\omega_1, \ldots, \omega_n) \in \{\text{head, tail}\}^n = \Omega$$
Last time... Discrete Distributions

- **Bernoulli distribution: Ber(p)**

  \[\Omega = \{\text{head, tail}\} \quad X(\text{head}) = 1, \quad X(\text{tail}) = 0.\]

  \[P(X = a) = P(\omega : X(\omega) = a) = \begin{cases} 
  p, & \text{for } a = 1 \\
  1 - p, & \text{for } a = 0 
\end{cases}\]

- **Binomial distribution: Bin(n,p)**

Suppose a coin with head prob. \(p\) is tossed \(n\) times. What is the probability of getting \(k\) heads and \(n-k\) tails?

\[\Omega = \{\text{possible } n \text{ long head/tail series}\}, \quad |\Omega| = 2^n\]

\(K(\omega) = \text{number of heads in } \omega = (\omega_1, \ldots, \omega_n) \in \{\text{head, tail}\}^n = \Omega\)

\[P(K = k) = P(\omega : K(\omega) = k) = \sum_{\omega : K(\omega) = k} p^k (1-p)^{n-k} = \binom{n}{k} p^k (1-p)^{n-k}\]
Last time... Conditional Probability

\[ P(X|Y) = \text{Fraction of worlds in which } X \text{ event is true given } Y \text{ event is true.} \]

\[ P(X|Y) = \frac{P(X, Y)}{P(Y)} \]
Last time… **Conditional Probability**

$P(X|Y) = \text{Fraction of worlds in which X event is true given Y event is true.}$

$$P(X|Y) = \frac{P(X, Y)}{P(Y)}$$

$$P(\text{flu|headache}) = \frac{P(\text{flu, headache})}{P(\text{headache})} = \frac{1/80}{1/80 + 7/80}$$

<table>
<thead>
<tr>
<th></th>
<th>Flu</th>
<th>No Flu</th>
</tr>
</thead>
<tbody>
<tr>
<td>Headache</td>
<td>1/80</td>
<td>7/80</td>
</tr>
<tr>
<td>No Headache</td>
<td>1/80</td>
<td>71/80</td>
</tr>
</tbody>
</table>
Independence

Independent random variables:

\[ P(X, Y) = P(X)P(Y) \]
\[ P(X|Y) = P(X) \]

Y and X don’t contain information about each other. Observing Y doesn’t help predicting X. Observing X doesn’t help predicting Y.

Examples:
Independent: Winning on roulette this week and next week.
Dependent: Russian roulette
Dependent / Independent

Independent X,Y

Dependent X,Y
Conditionally independent:

\[ P(X, Y | Z) = P(X | Z) P(Y | Z) \]

Knowing Z makes X and Y independent

Examples:
Dependent: shoe size of children and reading skills
Conditionally independent: shoe size of children and reading skills given age

Stork deliver babies:
Highly statistically significant correlation exists between stork populations and human birth rates across Europe.
London taxi drivers: A survey has pointed out a positive and significant correlation between the number of accidents and wearing coats. They concluded that coats could hinder movements of drivers and be the cause of accidents. A new law was prepared to prohibit drivers from wearing coats when driving.

Finally, another study pointed out that people wear coats when it rains...
Correlation ≠ Causation

Number people who drowned by falling into a swimming-pool correlates with Number of films Nicolas Cage appeared in

Correlation: 0.666004

http://www.tylervigen.com
Conditional Independence

Formally: X is **conditionally independent** of Y given Z

\[ P(X, Y|Z) = P(X|Z)P(Y|Z) \]

\[ P(\text{Accidents, Coats} | \text{Rain}) = P(\text{Accidents} | \text{Rain})P(\text{Coats} | \text{Rain}) \]
Conditional Independence

Formally: $X$ is \textbf{conditionally independent} of $Y$ given $Z$

\[ P(X, Y | Z) = P(X | Z)P(Y | Z) \]

\[ P(\text{Accidents, Coats} | \text{Rain}) = P(\text{Accidents} | \text{Rain})P(\text{Coats} | \text{Rain}) \]

Equivalent to:

\[ (\forall x, y, z)P(X = x | Y = y, Z = z) = P(X = x | Z = z) \]
Conditional Independence

Formally: $X$ is **conditionally independent** of $Y$ given $Z$

$$P(X, Y|Z) = P(X|Z)P(Y|Z)$$

$P(\text{Accidents, Coats}|\text{Rain}) = P(\text{Accidents}|\text{Rain})P(\text{Coats}|\text{Rain})$

Equivalent to:

$$(\forall x, y, z)P(X = x|Y = y, Z = z) = P(X = x|Z = z)$$

$$P(\text{Thunder}|\text{Rain, Lightning}) = P(\text{Thunder}|\text{Lightning})$$

**Note:** does NOT mean Thunder is independent of Rain

**But** given Lightning knowing Rain doesn’t give more info about Thunder
Conditional vs. Marginal Independence

• C calls A and B separately and tells them a number \( n \in \{1, \ldots, 10\} \)
• Due to noise in the phone, A and B each imperfectly (and independently) draw a conclusion about what the number was.
• A thinks the number was \( n_a \) and B thinks it was \( n_b \).
• Are \( n_a \) and \( n_b \) marginally independent?
  - No, we expect e.g. \( P(n_a = 1|n_b = 1) > P(n_a = 1) \)
• Are \( n_a \) and \( n_b \) conditionally independent given \( n \)?
  - Yes, because if we know the true number, the outcomes \( n_a \) and \( n_b \) are purely determined by the noise in each phone.

\[
P(n_a = 1|n_b = 1, n=2) = P(n_a = 1|n=2)
\]
Parameter estimation: MLE, MAP
I have a coin, if I flip it, what’s the probability that it will fall with the head up?
Flipping a Coin

I have a coin, if I flip it, what’s the probability that it will fall with the head up?

Let us flip it a few times to estimate the probability:
I have a coin, if I flip it, what’s the probability that it will fall with the head up?

Let us flip it a few times to estimate the probability:
I have a coin, if I flip it, what’s the probability that it will fall with the head up?

Let us flip it a few times to estimate the probability:

The estimated probability is: $\frac{3}{5}$  “Frequency of heads”
Flipping a Coin

The estimated probability is: \( \frac{3}{5} \) “Frequency of heads”

Questions:
(1) Why frequency of heads???
(2) How good is this estimation???
(3) Why is this a machine learning problem???

We are going to answer these questions
Why frequency of heads???

• Frequency of heads is exactly the maximum likelihood estimator for this problem

• MLE has nice properties (interpretation, statistical guarantees, simple)
Maximum Likelihood Estimation
MLE for Bernoulli distribution

Data, $D = \{X_i\}_{i=1}^{n}$, $X_i \in \{H, T\}$

$P(Heads) = \theta$, $P(Tails) = 1-\theta$
MLE for Bernoulli distribution

Data, $D = \{X_i\}_{i=1}^n$, $X_i \in \{H, T\}$

$P(Heads) = \theta, \; P(Tails) = 1-\theta$

Flips are i.i.d.:
MLE for Bernoulli distribution

Data, $D = \{X_i\}_{i=1}^n, \ X_i \in \{H, T\}$

$P(Heads) = \theta, \ P(Tails) = 1-\theta$

Flips are i.i.d.:
- Independent events
  - Identically distributed according to Bernoulli distribution
MLE for Bernoulli distribution

Data, $D =$

$$D = \{X_i\}_{i=1}^n, \; X_i \in \{H, T\}$$

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

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

**MLE:** Choose $\theta$ that maximizes the probability of observed data
Maximum Likelihood Estimation

MLE: Choose $\theta$ that maximizes the probability of observed data

$$\hat{\theta}_{MLE} = \arg \max_{\theta} P(D|\theta)$$
Maximum Likelihood Estimation

MLE: Choose \( \theta \) that maximizes the probability of observed data

\[
\hat{\theta}_{MLE} = \arg \max_{\theta} P(D|\theta) = \arg \max_{\theta} \prod_{i=1}^{n} P(X_i|\theta)
\]

independent draws
Maximum Likelihood Estimation

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

\[ = \arg \max_{\theta} \prod_{i=1}^{n} P(X_i|\theta) \]

\[ = \arg \max_{\theta} \prod_{i:X_i=H} \theta \prod_{i:X_i=T} (1 - \theta) \]

MLE: Choose \( \theta \) that maximizes the probability of observed data.

Independent draws identically distributed
Maximum Likelihood Estimation

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

\[ = \arg \max_{\theta} \prod_{i=1}^{n} P(X_i|\theta) \]

\[ = \arg \max_{\theta} \prod_{i:X_i=H} \theta \prod_{i:X_i=T} (1 - \theta) \]

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

MLE: Choose \( \theta \) that maximizes the probability of observed data

- Independent draws
- Identically distributed
Maximum Likelihood Estimation

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

\[ = \arg \max_{\theta} \prod_{i=1}^{n} P(X_i|\theta) \]

\[ = \arg \max_{\theta} \prod_{i:X_i=H} \theta \prod_{i:X_i=T} (1 - \theta) \]

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

MLE: Choose \( \theta \) that maximizes the probability of observed data

Independent draws

Identically distributed
Maximum Likelihood Estimation

MLE: Choose $\theta$ that maximizes the probability of observed data

$$\hat{\theta}_{MLE} = \arg \max_\theta P(D|\theta)$$

$$= \arg \max_\theta \theta^\alpha H (1 - \theta)^\alpha T$$

$$J(\theta)$$
MLE: Choose $\theta$ that maximizes the probability of observed data

$$\hat{\theta}_{MLE} = \arg \max_{\theta} P(D|\theta)$$

$$= \arg \max_{\theta} \theta^{\alpha_H} (1 - \theta)^{\alpha_T} J(\theta)$$

$$\frac{\partial J(\theta)}{\partial \theta} = \alpha_H \theta^{\alpha_H-1} (1 - \theta)^{\alpha_T} - \alpha_T \theta^{\alpha_H} (1 - \theta)^{\alpha_T-1} \bigg|_{\theta = \hat{\theta}_{MLE}} = 0$$
Maximum Likelihood Estimation

MLE: Choose $\theta$ that maximizes the probability of observed data

$$\hat{\theta}_{MLE} = \arg \max_{\theta} P(D|\theta)$$

$$= \arg \max_{\theta} \theta^{\alpha_H} (1 - \theta)^{\alpha_T}$$

$$J(\theta)$$

$$\frac{\partial J(\theta)}{\partial \theta} = \alpha_H \theta^{\alpha_H - 1} (1 - \theta)^{\alpha_T} - \alpha_T \theta^{\alpha_H} (1 - \theta)^{\alpha_T - 1} \bigg|_{\theta = \hat{\theta}_{MLE}} = 0$$

$$\alpha_H (1 - \theta) - \alpha_T \theta \bigg|_{\theta = \hat{\theta}_{MLE}} = 0$$
Question (2)

• How good is this MLE estimation???

\[ \hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T} \]
How many flips do I need?

I flipped the coins 5 times: 3 heads, 2 tails

$$\hat{\theta}_{MLE} = \frac{3}{5}$$

What if I flipped 30 heads and 20 tails?

$$\hat{\theta}_{MLE} = \frac{30}{50}$$

- Which estimator should we trust more?
- The more the merrier???
Let $\theta^*$ be the true parameter.

For $n = \alpha_H + \alpha_T$, and $\hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T}$

For any $\epsilon > 0$:

**Hoeffding’s inequality:**

$$P(|\hat{\theta} - \theta^*| \geq \epsilon) \leq 2e^{-2n\epsilon^2}$$
Simple bound

For $n = \alpha H + \alpha T$, and Hoeffding’s inequality:

For any $\varepsilon > 0$:

Let $\theta^*$ be the true parameter.

I want to know the coin parameter $\theta$, within $\varepsilon = 0.1$ error with probability at least $1-\delta = 0.95$.

How many flips do I need?

$$P(\left| \hat{\theta} - \theta^* \right| \geq \varepsilon) \leq 2e^{-2n\varepsilon^2}$$

Sample complexity:

$$n \geq \frac{\ln(2/\delta)}{2\varepsilon^2}$$
Question (3)

Why is this a machine learning problem???

• improve their **performance** (accuracy of the predicted prob.)
• at some **task** (predicting the probability of heads)
• with **experience** (the more coins we flip the better we are)
What about continuous features?

Let us try Gaussians...

\[ p(x | \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right) = N_x(\mu, \sigma) \]
MLE for Gaussian mean and variance

Choose $\theta= (\mu, \sigma^2)$ that maximizes the probability of observed data

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

$$= \arg \max_{\theta} \prod_{i=1}^{n} P(X_i \mid \theta)$$

$$= \arg \max_{\theta} \prod_{i=1}^{n} \frac{1}{2\sigma^2} e^{-\frac{(X_i-\mu)^2}{2\sigma^2}}$$

$$= \arg \max_{\theta= (\mu, \sigma^2)} \frac{1}{2\sigma^2} e^{-\sum_{i=1}^{n} \frac{(X_i-\mu)^2}{2\sigma^2}}$$

\[ J(\theta) \]
MLE for Gaussian mean and variance

\[
\hat{\mu}_{MLE} = \frac{1}{n} \sum_{i=1}^{n} x_i
\]

\[
\hat{\sigma}^2_{MLE} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2
\]

**Note:** MLE for the variance of a Gaussian is **biased**
[Expected result of estimation is not the true parameter!]

Unbiased variance estimator:
\[
\hat{\sigma}^2_{unbiased} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \hat{\mu})^2
\]
Probably Approximate Correct (PAC) Learning

I want to know the coin parameter \( \theta \), within \( \epsilon = 0.1 \) error with probability at least \( 1-\delta = 0.95 \).

How many flips do I need?

\[
P(|\hat{\theta} - \theta^*| \geq \epsilon) \leq 2e^{-2n\epsilon^2}
\]

Sample complexity:

\[
n \geq \frac{\ln(2/\delta)}{2\epsilon^2}
\]
What about prior knowledge? (MAP Estimation)
What about prior knowledge?

We know the coin is “close” to 50-50. What can we do now?

The Bayesian way...
What about prior knowledge?

We know the coin is “close” to 50-50. What can we do now?

The Bayesian way...

Rather than estimating a single $\theta$, we obtain a distribution over possible values of $\theta$. 

![Diagram showing the transition from prior to posterior distribution for $\theta$.](image-url)
Prior distribution

- What prior? What distribution do we want for a prior?
  - Represents expert knowledge (philosophical approach)
  - Simple posterior form (engineer’s approach)

- Uninformative priors:
  - Uniform distribution

- Conjugate priors:
  - Closed-form representation of posterior
  - $P(\theta)$ and $P(\theta|D)$ have the same form
In order to proceed we will need:

Bayes Rule

Chain Rule & Bayes Rule

Chain rule:
\[ P(X, Y) = P(X|Y)P(Y) = P(Y|X)P(X) \]

Bayes rule:
\[ P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)} \]

Bayes rule is important for reverse conditioning.
Bayesian Learning

• Use Bayes rule:

\[ P(\theta \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \theta)P(\theta)}{P(\mathcal{D})} \]

• Or equivalently:

\[ P(\theta \mid \mathcal{D}) \propto P(\mathcal{D} \mid \theta)P(\theta) \]

posterior likelihood prior

MAP estimation for Binomial distribution

Coin flip problem

Likelihood is Binomial

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

If the prior is Beta distribution,

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

\[ \Rightarrow \text{posterior is Beta distribution} \]

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

\[ P(\theta) \text{ and } P(\theta \mid D) \text{ have the same form! [Conjugate prior]} \]

\[ \hat{\theta}_{\text{MAP}} = \arg \max_\theta P(\theta \mid D) = \arg \max_\theta P(D \mid \theta)P(\theta) = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \alpha_T + \beta_T - 2} \]
Beta distribution

More concentrated as values of $\alpha$, $\beta$ increase
Beta conjugate prior

\[ P(\theta) \sim \text{Beta}(\beta_H, \beta_T) \quad \quad \quad P(\theta|D) \sim \text{Beta}(\beta_H + \alpha_H, \beta_T + \alpha_T) \]

As \( n = \alpha_H + \alpha_T \) increases

As we get more samples, effect of prior is “washed out”
**C3PO:** Sir, the possibility of successfully navigating an asteroid field is approximately 3,720 to 1!

**Han:** Never tell me the odds! \[ P(\theta) \sim Beta(\beta_H, \beta_T) \quad P(\theta|D) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T) \]

[Graphs showing C3PO's data backed beliefs, belief that Han will succeed, and posterior probability of success]

https://www.countbayesie.com/blog/2015/2/18/hans-solo-and-bayesian-priors
MLE vs. MAP

- Maximum Likelihood estimation (MLE)
  Choose value that maximizes the probability of observed data
  \[
  \hat{\theta}_{MLE} = \arg \max_{\theta} P(D | \theta)
  \]
MLE vs. MAP

- **Maximum Likelihood estimation (MLE)**
  Choose value that maximizes the probability of observed data
  \[ \hat{\theta}_{MLE} = \arg \max_{\theta} P(D|\theta) \]

- **Maximum a posteriori (MAP) estimation**
  Choose value that is most probable given observed data and prior belief
  \[ \hat{\theta}_{MAP} = \arg \max_{\theta} P(\theta|D) \]
  \[ = \arg \max_{\theta} P(D|\theta) P(\theta) \]

**When is MAP same as MLE?**
From Binomial to Multinomial

**Example:** 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.

Bayesians vs. Frequentists

You are no good when sample is small

You give a different answer for different priors
Next Class:

Naïve Bayes Classifier