Introduction to statistical models

- Learning from training data usually means estimating the parameters of the statistical model
- Estimation usually carried out via machine learning
- Supervised learning
  - Training data consists of the inputs and respective outputs (labels)
  - Labels are usually created via expert annotation (expensive)
  - Difficult to annotate when predicting more complex outputs
- Unsupervised learning
  - Training data just consists of inputs. No labels.
  - One example of such an algorithm: Expectation Maximization
What is MLE?

- Given
  - A sample $X=\{X_1, ..., X_n\}$
  - A vector of parameters $\theta$

- We define
  - Likelihood of the data: $P(X | \theta)$
  - Log-likelihood of the data: $L(\theta) = \log P(X | \theta)$

- Given $X$, find
  $$\theta_{ML} = \arg \max_{\theta \in \Theta} L(\Theta)$$

Model Space

- The choice of the model space is plentiful but not unlimited.
- There is a bit of “art” in selecting the appropriate model space.
- Typically the model space is assumed to be a linear combination of known probability distribution functions.
Example (I)

- Example: tossing a coin
  - Head, tail, head, head, tail, tail, ...
- Bernoulli distribution \( B(p) \):
  - probability distribution of a random variable which takes value 1 with success probability \( p \) and value 0 with failure probability \( 1 - p \)
  - Probability distribution

\[
P(X = x) = p^x (1 - p)^{1-x}
\]

Two Important Facts

- If \( X_1, \ldots, X_n \) are independent and identically distributed then
  \[
P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i)
\]
- The log function is monotonically increasing
  \Rightarrow \text{a function } f(x) \geq 0 \text{, achieves a maximum at } x_1 \text{, then log(f(x)) also achieves a maximum at } x_1.
- Log calculations allows to split a product into summation
  \[
  \log(P(X_1, \ldots, X_n)) = \sum_{i=1}^{n} \log(P(X_i))
  \]
Example (II)

• Suppose we have the following observation of a coin toss
  - \( X[] = \{0,0,0,1,0,0,0,0,0,1\} \)
  - What is the MLE of \( p \) given the observed data?

\[
P(X_1, \ldots X_n) = \prod_{i=1}^{n} p^{\delta(X_i, 1)}(1 - p)^{\delta(X_i, 0)}
\]

\[
= p^2(1 - p)^8
\]

- To determine value of \( p \) that fits best the observed data, differentiate with respect to \( p \) and find where the resulting expression is 0
- Slightly simpler calculation when using \( \log(P(X_1, \ldots X_n)) \)

Example (III)

\[
\frac{d}{dp} \left[ \log(P(X_1, \ldots X_{10})) \right] = 0
\]

\[
\frac{d}{dp} \left[ \frac{2 \log p + 8 \log (1 - p)}{} \right] = 0
\]

\[
\frac{2}{p} - \frac{8}{1 - p} = 0
\]

\[
p = 0.2
\]
Other examples

Suppose the following are marks in a course
55.5, 67, 87, 48, 63
Marks typically follow a Normal distribution whose density function is

\[ N(\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2\sigma}(x-\mu)^2} \]

Now, we want to find the best \( \mu, \sigma \) such that

\[ \arg\max_{\mu, \sigma} p(\text{Data}|\mu, \sigma) \]

Other examples

- Suppose we have data about heights of people (in cm)
  - 185, 140, 134, 150, 170
- Heights follow a normal (log normal) distribution but men and women differ in their overall distribution. This suggests a mixture of two distributions

\[ \pi_1 N(\mu_1, \sigma_1) + \pi_2 N(\mu_2, \sigma_2) \]
Maximum Likelihood Estimation

- We have reduced the problem of selecting the best model to that of selecting the best parameter.
- We want to select a parameter \( p \) which will maximize the probability that the data was generated from the model with the parameter \( p \) plugged-in.
- The parameter \( p \) is called the maximum likelihood estimator.
- The maximum of the function can be obtained by setting the derivative of the function \( ==0 \) and solving for \( p \).

What is MLE?

- Given
  - A sample \( X = \{X_1, ..., X_n\} \)
  - A vector of parameters \( \theta \)

- We define
  - Likelihood of the data: \( P(X | \theta) \)
  - Log-likelihood of the data: \( L(\theta) = \log P(X | \theta) \)

- Given \( X \), find

\[
\theta_{ML} = \arg \max_{\theta \in \Omega} L(\Theta)
\]
Basic setting in EM

- X is a set of data points: observed data
- Θ is a parameter vector.
- EM is a method to find θ_{ML} where

$$\theta_{ML} = \arg\max_{\theta \in \Omega} L(\Theta) = \arg\max_{\theta \in \Omega} \log P(X | \theta)$$

- In many cases: Calculating P(X,Y|θ) is where Y is “hidden” data (or “missing” data).

The basic EM strategy

- Z = (X, Y)
  - Z: complete data (“augmented data”)
  - X: observed data (“incomplete” data)
  - Y: hidden data (“missing” data)
- An example: three pegs that influence the path of a marble
  - Modeled through 3 Bernoulli random variables
  - Θ = (p_0, p_1, p_2) corresponding to the probabilities that the marble will go to the left or right of peg i
  - X = (N_0, N_1, N_2) number of times a marble ended up in the corresponding cup
  - Y = (,,,) number of times path 0,1,2,3, was taken
An EM Example

- Assumption: you can not observe which path a marble takes, only the final outcome
  - Y is ‘hidden’
  - X is observed
- Given a vector X determine \( \Theta \)
- Optimization criterion: sum over all settings of the hidden variable Y
  \[
P(X = x) = \sum_{y \in Y} P(X = x, Y = y; \theta)
\]
- Or for a vector of observed data
  \[
P(X) = \prod_{j=1}^{[X]} \sum_{y \in Y} P(X = x_j, Y = y; \theta)
\]
  \[
q(x, y, \theta^{(i)}) = \frac{\binom{X}{x} P(Y = y|X = x; \theta^{(i)})}{\sum_{y \in Y} \binom{X}{x} P(x, y; \theta^{(i)})}
\]
  with \( \sum_{(x,y) \in X \times Y} q(x, y) = 1 \)
- M-step calculate \( \Theta^{i+1} \) such that the probability of the joint distribution computed in the E-step is maximized
  
  \[
\theta^{(i+1)} = \arg \max_{\theta'} \sum_{(x,y,x,y)} q(x, y, \theta^{(i)}) \log P(X = x, Y = y, \theta')
\]

Expectation Maximization: iterative algorithm to calculate \( \Theta \) in an iterative manner, such that

\[
I(\theta^0) < I(\theta^1) < ... < I(\theta') < ....
\]

Two steps:
- E-step: compute the posterior probability of each possible hidden variable assignment \( y \in Y \) for each \( x \in X \) weighted by the relative frequency with which \( x \) occurs in \( X \)
  \[
q(x, y, \theta^{(i)}) = \frac{\binom{X}{x} P(Y = y|X = x; \theta^{(i)})}{\sum_{y \in Y} \binom{X}{x} P(x, y; \theta^{(i)})}
\]
  with \( \sum_{(x,y) \in X \times Y} q(x, y) = 1 \)
- M-step calculate \( \Theta^{i+1} \) such that the probability of the joint distribution computed in the E-step is maximized
  \[
\theta^{(i+1)} = \arg \max_{\theta'} \sum_{(x,y,x,y)} q(x, y, \theta^{(i)}) \log P(X = x, Y = y, \theta')
\]
• Compute steps:
  - Assuming \( X=\{N_a, N_b, N_c\} \) number of times marble ends in cups a, b, c calculate the probability for each different outcome that path \( y \in Y \) has been taken
  - Of all possible combination \( (x, y) \in X \times Y \) only 4 combinations matter, all other are 0:
    - \( P(Y=0 \mid X=a) = 1 \) => if marble is in cup a, it had to take path 0
    - \( P(Y=3 \mid X=c) = 1 \) => if marble is in cup c, it had to take path 3

\[
\Pr(1 \mid b; \theta^{(i)}) = \frac{(1 - p_0^{(i)})p_1^{(i)}}{(1 - p_0^{(i)})p_1^{(i)} + p_0^{(i)}(1 - p_2^{(i)})} \quad \Pr(2 \mid b; \theta^{(i)}) = \frac{p_0^{(i)}(1 - p_2^{(i)})}{(1 - p_0^{(i)})p_1^{(i)} + p_0^{(i)}(1 - p_2^{(i)})}
\]

**E-step**

• Since \( f \left( \frac{X}{N} \right) \approx \frac{N_X}{N} \)

<table>
<thead>
<tr>
<th>( X )</th>
<th>( Y )</th>
<th>( q(X=x, Y=y; \Theta^{(i)}) )</th>
<th>( \log \Pr(X, Y; \theta) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>0</td>
<td>( \frac{N_a}{N} )</td>
<td>( \log(1 - p_0) + \log(1 - p_1) )</td>
</tr>
<tr>
<td>b</td>
<td>1</td>
<td>( \Pr(1 \mid b; \theta^{(i)}) \cdot \frac{N_b}{N} )</td>
<td>( \log(1 - p_0) + \log p_1 )</td>
</tr>
<tr>
<td>b</td>
<td>2</td>
<td>( \Pr(2 \mid b; \theta^{(i)}) \cdot \frac{N_b}{N} )</td>
<td>( \log p_0 + \log(1 - p_2) )</td>
</tr>
<tr>
<td>c</td>
<td>3</td>
<td>( \frac{N_c}{N} )</td>
<td>( \log p_0 + \log p_2 )</td>
</tr>
</tbody>
</table>

• Multiply across each row and add the four resulting values
• Calculate derivative and set to zero
  - Done by treating each parameter separately
EM clustering

- Clustering: problem of estimating missing data
  - The missing data are the cluster labels

- Notice the similarity between EM for Normal mixtures and K-means.
  - The expectation step is the assignment.
  - The maximization step is the update of centres.