Background: Maximum Likelihood Estimation

CS534
Parameter Estimation

• Given observation of some random variable(s), assuming the variable(s) follow some fixed distribution, how to estimate the parameters?

• Example:
  – coin tosses (Bernoulli distribution, parameter: \( p \) – probability of head)
  – Dice or grades (Discrete distribution, parameters: \( p_i \) for \( i = 1, \ldots, k - 1 \) with \( k \) categories)
  – Height and weight of a person (continuous variable, parameters depends on the distribution, for example for Gaussian Distribution, the parameters are \( \mu \), the mean and \( \Sigma \), the covariance)
Maximum Likelihood Principle

• We will use a general term $M$ to denote the model (parameters) we try to estimate

• We will use a general term $D$ to denote the data that we observe
  – $D$ is a set of examples, let $D_m$ denote the $m$-th example in $D$

• Assuming a fixed model $M$, what is the probability of observing $D$?
  – This can be written as $P(D; M)$

• The maximum likelihood principle seeks to find $M$ that maximizes this probability

• This is called the likelihood function
  – $L(M) = P(D; M) = \prod_{m=1}^{n} P(D_m; M)$ --- here we are making the IID assumption, that is examples in $D$ are independently and identically distributed, thus we can use the product

• It is often more convenient to work with the log of $L(M)$
  – $l(M) = \log L(M) = \sum_{m=1}^{n} \log P(D_m; M)$
Example: Coin Toss

• You are given a coin, and want to decide $p$ - the probability of head of this coin
• You toss it for 500 times, and observe heads 242 times
• What is your estimate of $p$?

$$p = \frac{242}{500} = 0.484$$

• But why? – this is doing maximum likelihood estimation
• In other words, $p = 0.484$ leads to the highest probability of observing 242 heads out of 500 tosses
• Now let’s go over the math to convince ourselves
Example Cont.

- We have $D = \{x_1, \ldots, x_m\}$, where $x_m$ is the outcome of the $m$-th coin toss,
  - 1 means head, and 0 means tail
- Let's write down the likelihood function: $L(p) = P(D; p) = \prod_{m=1}^{n} P(x_m; p)$
- For binary variable, we have a nice compact form to write down its probability mass function $P(x_m; p) = p^{x_m} (1 - p)^{1-x_m}$
- So we have

\[
L(p) = \prod_{m=1}^{n} p^{x_m} (1 - p)^{1-x_m}
\]

- Take the log:

\[
l(p) = \log L(p) = \sum_{m=1}^{n} \log p^{x_m} (1 - p)^{1-x_m}
\]

\[
= \sum_{m=1}^{n} x_m \log p + \sum_{m=1}^{n} (1 - x_m) \log(1 - p) = n_1 \log p + n_0 \log(1 - p)
\]

\[
n_1 = \sum_{m=1}^{n} x_m \text{ and } n_0 = n - n_1
\]
Maximizing the likelihood

- We take derivative of $l(p)$:
  \[
  \frac{dl(p)}{dp} = \frac{n_1}{p} - \frac{n_0}{1-p}
  \]
- Setting it to zero:
  \[
  \frac{n_1}{p} = \frac{n_0}{1-p}
  \]
- Solving this leads to:
  \[
  p = \frac{n_1}{n}
  \]