CS 6140: Machine Learning **Spring 2018**
College of Computer and Information Science
Northeastern University
January, 22
Instructor: Bilal Ahmed **Scribe: Bilal Ahmed**

Parameter Estimation Methods - I[1]

1 A Coin Tossing Experiment

Consider that you are given a coin and asked to estimate the probability that the coin landing heads. What is a good way of doing this? Taking a practical approach we can toss the coin multiple times and record the number of times it turns up heads. Then our estimate of the probability that the coin will land heads will be given by the ratio of the observed number of heads to the number of trials. This is a very intuitive and straight-forward way to give an answer that makes sense, however is there an underlying mathematical principle that supports this answer? In this lecture we will formalize this line of reasoning to develop a framework for fitting parametric models to observed data.

1.1 A Parametric Model for Coin Tossing

Lets begin by developing a parametric model for a single coin toss. Let μ be the probability of the coin landing heads i.e., $P(x = 1)$ where $x \in \{0, 1\}$ is a binary random variable (0 represents tails while 1 represents heads). For a binary random variable, a single outcome/trial is modeled by the Bernoulli distribution, which takes the form:

$$Bern(x|\mu) = \mu^x \ (1 - \mu)^{1-x} \quad 0 \leq \mu \leq 1 \tag{1}$$

The task at hand is then to estimate μ from given data: the outcome of a number of coin tosses (our initial experiment). Once we are able to obtain an estimate of μ that agrees with our observed data, we would have fitted our parametric model to data.

1.2 Data Likelihood

We will treat each coin-toss (trial) as an independent observation, i.e., the outcome of any given coin-toss is not influenced by past observations and consequently a given coin-toss does not influence any future tosses. However, each observation is generated by the same underlying $Bern(x|\mu)$ distribution (we do not know what the value of μ) is, yet). This assumption is known as the IID assumption: all observations are independent and identically distributed. For data ($\mathcal{D}$) comprising of N coin tosses $\mathcal{D} = \{x_1, x_2, \ldots, x_N\}$, the IID assumption allows us to write the joint probability of the observations as:

[1] These lecture notes are intended for in-class use only.

$$P(\mathcal{D}|\mu) = \prod_{i=1}^{N} \mu^{x_i} \ (1-\mu)^{1-x_i} \quad 0 \leq \mu \leq 1 \tag{2}$$

We can interpret Equation-2 as the probability or likelihood of observing the data, given the model parameter(s). An important point to note here is that the data is fixed or known, while the model parameters are unknown, therefore we will treat Equation-2 as a function of the unknown parameters. We can estimate the parameters by maximizing the likelihood of data i.e., select the value of the model parameter such that the probability of the data being generated by the model is maximized. The resulting parameter estimates are known as maximum-likelihood estimates (MLE). Remember, even though Equation-2 denotes the probability of $\mathcal{D}$, for the purpose of parameter estimation we treat it as a function of the model parameters ($L(\mu; \mathcal{D})$).

1.3 Maximizing Likelihood

The likelihood function can be maximized with respect to the model parameters by differentiating with respect to the model parameters and equating to zero. As we are going to be working with the product of probabilities of independent events it is more convenient to work with the log-likelihood (LL). One of the advantages of working in the log domain is that we avoid numerical underflow that can arise as a result of multiplying N probability values.

$$\ln L(\mu; \mathcal{D}) = \sum_{i=1}^{N} \{x_i \ln(\mu) + (1-x_i) \ln(1-\mu)\} \tag{3}$$

Since, $x_i \in \{0, 1\}$ we can see that $\sum_{i=1}^{N} \{x_i\}$ is the number of heads that we observe in our training data, which we will denote by N_H, and similarly $\sum_{i=1}^{N} \{1 - x_i\}$ is the number of tails denoted by N_T. The likelihood is then given as:

$$LL(\mu; \mathcal{D}) = N_H \ln(\mu) + N_T \ln(1-\mu)$$

Now to find μ^*:

$$\mu^* = \arg\max_{\mu} LL(\mu; \mathcal{D})$$

which can be obtained as:

$$\begin{aligned}
\frac{\delta}{\delta\mu} LL(\mu; \mathcal{D}) &= N_H \frac{\delta}{\delta\mu} \ln(\mu) + N_T \frac{\delta}{\delta\mu} \ln(1-\mu) \\
&= N_H \frac{1}{\mu} + N_T \frac{1}{1-\mu} \frac{\delta}{\delta\mu}(1-\mu) \\
&= N_H \frac{1}{\mu} + N_T \frac{1}{1-\mu}(-1)
\end{aligned}$$

equate the above equal to zero to obtain μ^*

$$\begin{aligned}\frac{\delta}{\delta\mu}LL(\mu;\mathcal{D}) &= N_H\frac{1}{\mu} - N_T\frac{1}{1-\mu} = 0 \\ &\Rightarrow N_H(1-\mu) - N_T\mu = 0 \\ &\Rightarrow \mu^* = \frac{N_H}{N} \quad (where, N = N_H + N_T)\end{aligned}$$

We can see that our intuitive estimate for the parameter μ is the same as the maximum likelihood estimate for μ for a given training dataset.

2 Summary

We have formulated a mathematical framework i.e., maximum likelihood for fitting parameteric models to data. To use this framework we first need to select an appropriate model for our data, e.g., for the coin tossing experiment we modeled the outcome of each trial using a Bernoulli disribution. Once we have an appropriate model we can then formulate the data likelihood and maximize it to find the optimal parameters for our data (in the MLE sense).