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 - II[1]

1 Coin Tossing: Redux

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? We saw in the previous lecture that this can be done using the maximum likelihood framework for estimating parameter values from data that comprises of repeated coin tossing trials and recording the number of heads we observe. One fundamental problem that we run into with MLE is that parameter estimtes tend to be myopic (short-sighted), especially when the data is limited or biased. For example, consider the pathological case where we have five trials and all trials resulted in heads. In this case, the MLE for our model parameter i.e., the probability of observing heads will be exactly one, thus ruling out the possibility of observing tails on any future trial. This will keep our model from generalizing to unseen data (i.e., our model will not be able to explain any trial that results in a tail). How can we alleviate this shortcoming in MLE?

1.1 Baye's Theorem

Baye's theorem plays a central role in pattern recognition and machine learning. In the context of parameter estimation, Baye's theorem states:

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)} \tag{1}$$

where, X represents the training data and θ are the model parameters. The denominator can be viewed as a normalization constant, which ensures that the resulting quantity is a valid probability distribution. We already know that $P(X|\theta)$ is the data likelihood (for our choice of the parametric model), which we maximize directly to obtain the maximum likelihood estimates for $\theta = \theta^{ML}$. $P(\theta)$ is known as the prior distribution, and encodes our prior beliefs, domain knowledge, or uncertainty about the model parameters.

1.2 Maximum A Posteriori (MAP) Estimates

Observe, that in Equation-1 we will end up with a posterior distribution on our model parameters θ. How do we get a point estimate from this distribution, as we did for maximum likelihood? We can take as our estimate the value of θ that maximizes the posterior distribution $P(\theta|X)$ i.e.,

$$\theta^{MAP} = \arg\max_{\theta} P(\theta|X)$$

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

θ^{MAP} is known as the *maximum a posteriori* (MAP) estimate of the model parameters.

1.3 Conjugate Priors

When the posterior and the prior distribution belong to the same family of distributions, then the posterior and the prior are refered to as being conjugate distributions, and the prior is called a conjugate prior. Depending on the form of the data likelihood we can select a conjugate prior, and we will be guaranteed the from of the posterior. For example, for a normal distribution, the conjugate prior is also given by a normal distribution. Whereas, selecting a conjugate prior makes the resulting calculations easier, it does not provide any insights about what the exact prior should be? If we were to select a normal prior, what should be its variance? it the prior is too wide (almost uniform) it will not provide any information to the estimation process such priors are known as uninformative priors, on the other hand if the distribution is very narrow (specific) the resultign MAP estimate can be far away from the true parameter values θ^*.

1.4 A Prior For Coin Tossing

Lets take a look at the form of the data likelihood for our coin tossing experiment to determine which distribution we want to use as a prior? Recall, that the likelihood is given as:

$$L(\mu; \mathcal{D}) = \prod_{i=1}^{N} \mu^{x}(1-\mu)^{1-x}$$

we can observe that the likelihood is made up of factors of the form $\mu^x(1-\mu)^{1-x}$, if we choose a prior that depends on the powers of μ and $(1-\mu)$ then we can ensure that the posterior will also have the same form (i.e., a conjugate prior). We can use a beta distribution as the conjugate prior in this case, which is given as:

$$Beta(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \tag{2}$$

where, $\Gamma(x)$ is known as the gamma function, and for positive integers it is given as $\Gamma(x+1) = x!$, and the coefficient $\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$ is a normalization constant. The mean of the Beta distribution is given as $E[\mu] = \frac{a}{a+b}$. Here, a and b are known as the hyperparameters of the model, and can be set to encode different levels of belief about the model parameters.

1.5 MAP For Coin Tossing

We can estimate the MAP values of the model paramters by maximizing the posterior, which we know is going to be a Beta distribution (because we are using a conjugate prior). The posterior can be obtained by multiplying the likelihood with the Beta prior.

$$\begin{aligned} P(\mu|\mathcal{D},a,b) &\propto \{\mu^{N_H}(1-\mu)^{N_T}\}\{\mu^{a-1}(1-\mu)^{b-1}\} \\ &\propto \mu^{N_H+a-1}(1-\mu)^{N_T+b-1} \end{aligned}$$

given the form of the posterior it can be seen that it is a Beta distribution given as:

$$P(\mu|\mathcal{D},a,b) = Beta(N_H+a, N_T+b)$$

we can see that by using the Beta prior we are essentially adding default counts (pseudo-counts) to our data. It is almost as if we started our experiment by assuming that we had already performed $a + b - 2$ trials, observing $a - 1$ heads and $b - 1$ tails. Note, that we now have a distribution over our model parameters, but not a point estimate. The point estimate we are interested is the one that maximizes the posterior.

$$\mu^{MAP} = \arg\max_{\mu} P(\mu|\mathcal{D}, a, b)$$

which is given by:

$$\mu^{MAP} = \frac{N_H + a - 1}{N_H + N_T + a + b - 2}$$

this can be obtained by maximizing the posterior or log-posterior (homework problem). From the MAP estimates we can see that even if in our dataset we do not observe any heads (or tails) the prior pseudo-counts would still enable our model to anticipate these outcomes in future trials.

2 Summary

We developed a Bayesian treatment of parameter estimation for fitting parametric models to data. Within this framework we can estimate the posterior distribution over our model parameters and also obtain point values by maximizing the posterior. 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. In addition to this, we also need to select an appropriate prior for our model parameters (most of the time we will be working with conjugate priors to make the task easy). Once we have made these two choices, we can then obtain the posterior using Baye's theorem and then maximize the posterior to obtain the MAP estimate for our model parameters.