Lecture Outline

Reading: Section 2.1

- Bernoulli Distribution
- ML Estimation for the Bernoulli Distributional Parameter
- MAP Estimation for the Bernoulli Distributional Parameter

2. Introduction

In Chapter 2, we will examine the problem of *density estimation*, that is given data $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ find the parameters, θ of the distribution of $p(\mathbf{x})$. We will estimate the distributional parameters by maximizing the likelihood (frequentist treatment), $p(\mathcal{D}|\theta)$ i.e. finding θ that best explains $\mathcal{D}$.

We have already seen in Lecture 5, that if the data, $\mathcal{D} = \{x_1, \ldots, x_N\}$ are assumed to be Gaussian distributed, the ML estimate for the distributional parameters is given by the sample mean and sample variance:

$$\mu_{\text{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n \tag{1}$$

and

$$\sigma^2_{\text{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{\text{ML}})^2. \tag{2}$$

By introducing priors on the parameters themselves, $p(\theta)$ we will estimate the distributional parameters by maximizing the posterior probability (Bayesian treatment), $p(\theta|\mathcal{D}) = p(\mathcal{D}|\theta)p(\theta)/p(\mathcal{D})$ i.e. finding θ that (possibly) even better explains $\mathcal{D}$.

https://en.wikipedia.org/wiki/Prior_probability

To make the math easier, we will let $p(\theta)$ have the same form as $p(\mathcal{D}|\theta)$, i.e. use a *conjugate prior*. We will solve the density estimation problem for discrete-valued data which has a Bernoulli distribution and continuous-valued data which has a Gaussian distribution.

2.1 Binary Variables (Bernoulli Distribution)

Consider a binary, discrete random variable $x \in \{0, 1\}$. The probability that $x = 1$ is governed by a parameter $0 \leq \mu \leq 1$ so that

$$p(x = 1|\mu) = \mu \tag{3}$$

and

$$p(x = 0|\mu) = 1 - \mu. \tag{4}$$

The probability distribution over x, known as the *Bernoulli* distribution can be written with a single, distributional parameter as

$$\begin{aligned} p(x|\mu) &= \text{Bern}(x|\mu) \\ &= \mu^x (1 - \mu)^{(1-x)}. \end{aligned} \tag{5}$$

We also note that

$$\mathbb{E}[x] = \mu \tag{6}$$
$$\text{var}[x] = \mu(1-\mu). \tag{7}$$

ML Estimation for the Bernoulli Distributional Parameter

Now, suppose we have a data set $\mathcal{D} = \{x_1, \ldots, x_N\}$ where $x_n \in \{0, 1\}$ which we wish to model with $\text{Bern}(x|\mu)$. Our approach is to estimate μ from $\mathcal{D}$ via maximization of the likelihood function.

Assuming the observations (data) are independent, the likelihood function is given by

$$p(\mathcal{D}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) \tag{8}$$
$$= \prod_{n=1}^{N} \mu^{x_n} (1-\mu)^{(1-x_n)}. \tag{9}$$

Equivalent to maximizing the likelihood function with respect to μ, we can maximize the log likelihood with respect to μ

$$\begin{aligned} \ln p(\mathcal{D}|\mu) &= \sum_{n=1}^{N} \ln p(x_n|\mu) \\ &= \sum_{n=1}^{N} \ln \left[\mu^{x_n} (1-\mu)^{(1-x_n)}\right] \\ &= \sum_{n=1}^{N} \left[x_n \ln \mu + (1-x_n) \ln(1-\mu)\right]. \end{aligned} \tag{10}$$

If we set the derivative of (10) with respect to μ to zero and solve for μ, we obtain the ML estimate

$$\begin{aligned} \mu_{\text{ML}} &= \frac{1}{N} \sum_{n=1}^{N} x_n \\ &= \frac{m}{N} \end{aligned} \tag{11}$$

where m is the number of observations (data points) where $x = 1$. Thus the ML setting for μ in the Bernoulli distribution is the sample mean or fraction of observations (data) having $x = 1$.

Recap: The likelihood for a given μ is

$$p(\mathcal{D}|\mu) = \mu^m (1-\mu)^{N-m} \tag{12}$$

and is maximized when $\mu = \mu_{\text{ML}}$ above.

Suppose we wish to compute the probability, $p(x = 1|\mathcal{D})$. We could use the predictive distribution

$$\begin{aligned} p(x = 1|\mathcal{D}) &= \mathbb{E}[x|\mathcal{D}] \\ &= \mu_{\text{ML}} \end{aligned} \tag{13}$$

since x is Bern, $\mathbb{E}[x] = \mu$.

Example: Suppose our data set is $\mathcal{D} = \{1, 1, 1\}$, then $\mu_{\text{ML}} = 3/3 = 1 = p(x = 1|\mathcal{D})$ and all our future predictions of x would be 1. This, of course, is probably unreasonable and highlights the over-fitting associated with ML.

MAP Estimation for the Bernoulli Distributional Parameter

To arrive at a more sensible estimate for μ we introduce a prior distribution, $p(\mu)$ over μ and solve for μ by maximizing the posterior, $p(\mu|\mathcal{D})$. In order to simplify the problem, we would like the functional form of the prior to be of the same form as the likelihood in (12), i.e. *conjugacy*. Thus, we assume a prior of the form

$$\begin{aligned} p(\mu) &= \text{Beta}(\mu|a,b) \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \end{aligned} \tag{14}$$

where $\Gamma(\cdot)$ is the gamma function[1] and the coefficient ensures normalization. The parameters a and b are called *hyperparameters* because they control the distribution of the parameter μ.

The mean and variance of a Beta distribution are given by

$$\mathbb{E}[\mu] = \frac{a}{a+b} \tag{15}$$

$$\text{var}[\mu] = \frac{ab}{(a+b)^2(a+b+1)}. \tag{16}$$

You can think of $a,\ b$ as the number of fictious observations where $x=1,\ x=0$ respectively.

From Bayes' theorem, the posterior is

$$p(\mu|\mathcal{D}) = \frac{p(\mathcal{D}|\mu)p(\mu)}{p(\mathcal{D})} \tag{17}$$

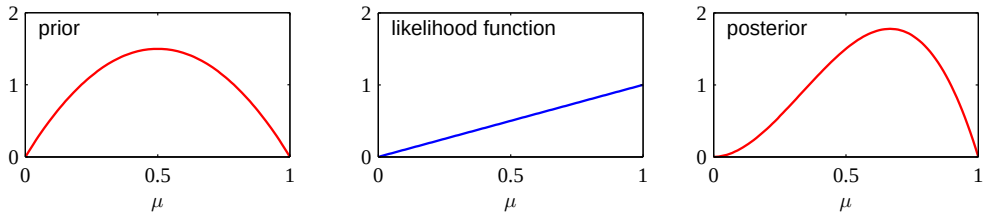
and keeping only the factors that depend on μ and omitting the normalization constant, we have

$$\begin{aligned} p(\mu|\mathcal{D}) &\propto p(\mathcal{D}|\mu)p(\mu) \\ &\propto \mu^m(1-\mu)^{N-m}\mu^{a-1}(1-\mu)^{b-1} \\ &\propto \mu^{m+a-1}(1-\mu)^{N-m+b-1}. \end{aligned} \tag{18}$$

Normalizing, we can show that with $l = N-m$, i.e. the number of observations of $x=0$

$$p(\mu|\mathcal{D}) = \frac{\Gamma(m+a+l+b)}{\Gamma(m+a)\Gamma(l+b)}\mu^{m+a-1}(1-\mu)^{l+b-1}. \tag{19}$$

We see that the effect of observing a data set of m observations of $x=1$ and l observations of $x=0$ has been to increase the value of a by m and the value of b by l, in going from the prior distribution $p(\mu)$ in (14) to the posterior distribution $p(\mu|\mathcal{D})$ (19).



(a) (b) (c)

[1] $\Gamma(x) \equiv \int_0^\infty u^{x-1}e^{-u}du$ and for $x \in \mathbb{Z}$, $\Gamma(x+1) = x\Gamma(x)$, $\Gamma(1)=1$. Thus $\Gamma(x) = (x-1)!$ for $x \in \mathbb{Z}$. $\Gamma(\cdot)$ is a smooth curve that connects $(x, \Gamma(x))$ points at the positive integers x.

If our goal is to predict as best we can, the outcome of the next trial, then we must evaluate the predictive distribution of x, given the observed data set $\mathcal{D}$:

$$\begin{aligned} p(x=1|\mathcal{D}) &= \int_0^1 p(x=1|\mu)p(\mu|\mathcal{D})d\mu \\ &= \int_0^1 \mu p(\mu|\mathcal{D})d\mu \\ &= \mathbb{E}[\mu|\mathcal{D}] \\ &= \frac{m+a}{m+a+l+b}. \end{aligned} \tag{20}$$

We interpret this result as the total fraction of observations (data) (both real and fictitious prior observations) that correspond to $x = 1$.

Example: Suppose our data set is $\mathcal{D} = \{1, 1, 1\}$ and the hyperparameters are $a = b = 10$. Then the probability that $p(x = 1|\mathcal{D}) = (3 + 10)/(3 + 10 + 0 + 10) = 0.57$. This, of course, is more reasonable than the prediction based on ML.

Note that in the limit of an infinitely large data set $m, l \to \infty$ and $p(x = 1|\mathcal{D})$ reduces to the ML result.

It is a very general property that the Bayesian (MAP) and ML results will agree in the limit of an infinitely large data set. For a finite data set, the posterior mean for μ $(m + a)/(m + a + l + b)$ lies between the prior mean $a/(a + b)$ and the ML estimate for μ $m/(m + l)$.

2.2 Multinomial Variables

Students should read this section on their own.