CSE 455/555 Spring 2013 Homework 7: Parametric Techniques

Jason J. Corso
Computer Science and Engineering
SUNY at Buffalo
jcorso@buffalo.edu
Solutions by Yingbo Zhou

This assignment does not need to be submitted and will not be graded, but students are advised to work through the problems to ensure they understand the material.

You are both allowed and encouraged to work in groups on this and other homework assignments in this class. These are challenging topics, and working together will both make them easier to decipher and help you ensure that you truly understand them.

---

1. **Maximum Likelihood of Binary/Multinomial Variable**
   Suppose we have a single binary variable $x \in \{0, 1\}$ with $x = 1$ denotes the 'heads' and $x = 0$ denotes the 'tails' of the outcome from flipping a coin. We do not make any assumption on the fairness of the coin, instead, we assume the probability of $x = 1$ will be denoted by a parameter μ so that $p(x = 1|\mu) = \mu$, where $0 \leq \mu \leq 1$.

   (a) Write down the probability distribution of x.
   Solution:
   $p(x = 0|\mu) = 1 - \mu$, so $p(x|\mu) = \mu^x(1 - \mu)^{1-x}$, this is known as the Bernoulli distribution.

   (b) Show that this is a proper probability distribution, *i.e.* the probability sum up to 1. What is the expectation and variance of this distribution?
   Solution:
   $\sum_{x\in\{0,1\}} p(x|\mu) = p(x = 0|\mu) + p(x = 1|\mu) = 1 - \mu + \mu = 1$
   $\mathbb{E}(x) = \sum_{x\in\{0,1\}} xp(x|\mu) = 0 \times p(x = 0|\mu) + 1 \times p(x = 1|\mu) = \mu$
   $\text{var}(x) = \{\mathbb{E}(x^2) - \mathbb{E}(x)^2\} = \sum_{x\in\{0,1\}} x^2 p(x|\mu) - \mu^2 = 0^2 \times p(x = 0|\mu) + 1^2 \times p(x = 1|\mu) - \mu^2 = \mu - \mu^2$

   (c) Now suppose we have a set of observed values $\mathbf{X} = \{x_1, x_2, \ldots, x_N\}$ of x. Write the likelihood function and estimate the maximum likelihood parameter μ_{ML}.
   Solution:
   $p(\mathbf{X}|\mu) = \prod_{i=1}^N p(x_i|\mu) = \prod_{i=1}^N \mu^{x_i}(1 - \mu)^{1-x_i}$
   The log-likelihood can be written as:
   $l(\mathbf{X}|\mu) = \sum_{i=1}^N \ln p(x_i|\mu) = \sum_{i=1}^N \{x_i \ln \mu + (1 - x_i) \ln(1 - \mu)\}$
   Take the derivative w.r.t. μ and set to zero we get:

$$\sum_{i=1}^N \{\frac{x_i}{\mu} - \frac{(1 - x_i)}{1 - \mu}\} = 0$$

$$(1 - \mu)\sum_{i=1}^N x_i = \mu \sum_{i=1}^N (1 - x_i)$$

$$\sum_{i=1}^N x_i - \mu \sum_{i=1}^N x_i = N\mu - \mu \sum_{i=1}^N x_i$$

$$\mu_{ML} = \frac{\sum_{i=1}^{N} x_i}{N}$$

if we observe m heads, then $\mu_{ML} = \frac{m}{N}$.

(d) Now suppose we are rolling a K-sided dice, in other words we have data $D = \{x_1, x_2, \ldots, x_N\}$ which can take on K values. Assume the generation of each value of $k \in K$ is determined by a parameter $\theta_k \geq 0$, so that the $p(x = k|\theta_k) = \theta_k$, and $\sum_k \theta_k = 1$. Write down the likelihood function.

Solution:

First we introduce a convenient representation called 1-of-K scheme, which the variables is represented by a K-dimensional binary vector that one of the element can take on value one. This is exactly the case for this question, since at each time we can only generate one sample taken on one particular value. Therefore, similar to the Bernoulli form we have the distribution shown as:

$$p(x|\theta) = \prod_{i=1}^{K} \theta_i^{x_i}$$

where $\sum_{i=1}^{K} x_i = 1$, $\theta_i \geq 0$ and $\sum_{i=1}^{K} \theta_i = 1$. Now we can write the likelihood function as:

$$p(D|\theta) = \prod_{n=1}^{N} \prod_{i=1}^{K} \theta_i^{x_{ni}} = \prod_{i=1}^{K} \theta_i^{\sum_n x_{ni}} = \prod_{i=1}^{K} \theta_i^{n_i}$$

where $n_i = \sum_n x_{ni}$ is the number of data that take on value i, and log-likelihood:

$$l(D|\theta) = \sum_{n=1}^{N} \sum_{i=1}^{K} x_{ni} \ln \theta_i$$

(e) Write the maximum likelihood solution of θ_k^{ML}.

Solution:

We have to maximize the likelihood w.r.t. constraint, so we have to introduce Lagrange multiplier, and the new objective function is:

$$F(\theta, \lambda) = \sum_{n=1}^{N} \sum_{i=1}^{K} x_{ni} \ln \theta_i + \lambda(\sum_{i=1}^{K} \theta_i - 1)$$

take the derivative w.r.t. θ_j and set to zero we get:

$$\begin{aligned} \frac{\partial F}{\partial \theta_j} &= \sum_{n=1}^{N} \frac{x_{nj}}{\theta_j} + \lambda = 0 \\ &\Rightarrow \frac{n_j}{\theta_j} = -\lambda \\ &\Rightarrow \theta_j = -\frac{n_j}{\lambda} \end{aligned}$$

substitute this back to our constraint $\sum_i \theta_i = 1$, we have

$$\begin{aligned} \sum_{i=1}^{K} \theta_i = -\sum_{i=1}^{K} \frac{n_i}{\lambda} = 1 \\ \Rightarrow -N = \lambda \end{aligned}$$

therefore we get $\theta_i = \frac{n_i}{N}$.

2. **Naive Bayes**

In naive Bayes, we assume that the presence of a particular feature of a class is unrelated to the presence of any other feature. For example, a fruit may be considered to be a watermelon if it is green, round, and more than 10 pounds. we will consider all these three features as independent to each other in naive Bayes.

Let our features $x_i, i \in [1\ d]$ be binary valued and have d dimensions, *i.e.* $x_i \in \{0, 1\}$ and our input feature vector $x = [x_1 x_2 \cdots x_d]^{\mathbf{T}}$. For each training sample, our target value $y \in \{0, 1\}$ is also a binary-valued variable. Then our model is parameterized by $\phi_{i|y=0} = p(x_i = 1|y = 0)$, $\phi_{i|y=1} = p(x_i = 1|y = 1)$, and $\phi_y = p(y = 1)$, and

$$
\begin{aligned}
p(y) &= (\phi_y)^y (1 - \phi_y)^{(1-y)} \\
p(x|y = 0) &= \prod_{i=1}^{d} p(x_i|y = 0) \\
&= \prod_{i=1}^{d} (\phi_{i|y=0})^{x_i} (1 - \phi_{i|y=0})^{(1-x_i)} \\
p(x|y = 1) &= \prod_{i=1}^{d} p(x_i|y = 1) \\
&= \prod_{i=1}^{d} (\phi_{i|y=1})^{x_i} (1 - \phi_{i|y=1})^{(1-x_i)}
\end{aligned}
$$

(a) Write down the joint log-likelihood function $l(\theta) = \log \prod_{n=1}^{N} p(x^{(n)}, y^{(n)}; \theta)$ in terms of the model parameters given above. $x^{(n)}$ means the nth data point, and θ represents all the parameters, *i.e.* $\{\phi_y, \phi_{i|y=0}, \phi_{i|y=1}, i = 1, \ldots, d\}$.
Solution:

$$
\begin{aligned}
l(\theta) &= \log \prod_{n=1}^{N} p(x^{(n)}, y^{(n)}; \theta) \\
&= \log \prod_{n=1}^{N} p(x^{(n)} \mid y = y^{(n)}; \theta) p(y^{(n)}; \theta) \\
&= \log \prod_{n=1}^{N} \left(\prod_{i=1}^{d} p(x_i^{(n)}|y^{(n)}; \theta) \right) p(y^{(n)}; \theta) \\
&= \sum_{n=1}^{N} \left(\log p(y^{(n)}; \theta) + \sum_{i=1}^{d} \log p(x_i^{(n)}|y^{(n)}; \theta) \right) \\
&= \sum_{n=1}^{N} \left\{ y^{(n)} \log \phi_y + (1 - y^{(n)} \log(1 - \phi_y) + \sum_{i=1}^{d} \left(x_i^{(n)} \log \phi_{i|y^{(n)}} + (1 - x_i^{(n)}) \log(1 - \phi_{i|y^{(n)}}) \right) \right\}
\end{aligned}
$$

(b) Estimate the paramters using maximum likelihood, *i.e.* find solutions for paramter ϕ_y, $\phi_{i|y=0}$ and $\phi_{i|y=1}$.
Solution:
Take derivative with respect to these 3 parameters, we get:

$$
\phi_y = \frac{\sum_{n=1}^{N} y^{(n)}}{N}
$$

$$\phi_{i|y=0} = \frac{\sum_{n=1}^{N}(1-y^{(n)})x_i^{(n)}}{\sum_{n=1}^{N}(1-y^{(n)})}$$

$$\phi_{i|y=1} = \frac{\sum_{n=1}^{N}y^{(n)}x_i^{(n)}}{\sum_{n=1}^{N}y^{(n)}}$$

(c) When a new sample point x comes, we make the prediction based on the most likely class estimate generated by our model. Show that the hypothesis returned by naive Bayes is linear, *i.e.* if $p(y=0|x)$ and $p(y=1|x)$ are the class probabilities returned by our model, show that there exist some α so that

$$p(y=1|x) \geq p(y=0|x) \text{if and only if } \alpha^{\mathbf{T}}\tilde{\mathbf{x}} \geq 0$$

where $\alpha = [\alpha_0\ \alpha_1\ \cdots\ \alpha_d]^{\mathbf{T}}$ and $\tilde{\mathbf{x}} = [1\ x_1\ x_2\ \cdots\ x_d]^{\mathbf{T}}$.
Solution:

$$\begin{aligned}
& p(y=1 \mid x) \geq p(y=0 \mid x) \\
\Leftrightarrow\ & p(x \mid y=0)p(y=0) \geq p(x \mid y=1)p(y=1) \\
\Leftrightarrow\ & (1-\phi_y)\prod_{i=1}^{d}(\phi_{i|y=0})^{x_i}(1-\phi_{i|y=0})^{(1-x_i)} \geq \phi_y \prod_{i=1}^{d}(\phi_{i|y=1})^{x_i}(1-\phi_{i|y=1})^{(1-x_i)} \\
\Leftrightarrow\ & \log(1-\phi_y) + \sum_{i=1}^{d} x_i \log\phi_{i|y=0} + \sum_{i=1}^{d}(1-x_i)\log(1-\phi_{i|y=0}) \\
\geq\ & \log\phi_y + \sum_{i=1}^{d} x_i \log\phi_{i|y=1} + \sum_{i=1}^{d}(1-x_i)\log(1-\phi_{i|y=1}) \\
\Leftrightarrow\ & \sum_{i=1}^{d} x_i \log\frac{\phi_{i|y=0}}{\phi_{i|y=1}}\frac{1-\phi_{i|y=1}}{1-\phi_{i|y=0}} + \log\left(\frac{1-\phi_y}{\phi_y}\left(\frac{1-\phi_{i|y=0}}{1-\phi_{i|y=1}}\right)^d\right) \geq 0 \\
\Leftrightarrow\ & \alpha^{\mathbf{T}}\tilde{\mathbf{x}} \geq 0
\end{aligned}$$

Where

$$\alpha_0 = \log\left(\frac{1-\phi_y}{\phi_y}\left(\frac{1-\phi_{i|y=0}}{1-\phi_{i|y=1}}\right)^d\right)$$

$$\alpha_i = \log\frac{\phi_{i|y=0}}{\phi_{i|y=1}}\frac{1-\phi_{i|y=1}}{1-\phi_{i|y=0}}$$

3. **Gaussian Distribution**

Please familiarize yourself with the maximum likelihood estimation for Gaussian distribution in the class notes and answer the following questions.

(a) What is the joint probability distribution $p(\mathbf{X};\mu,\sigma^2)$ of the samples?
Solution:

$$p(\mathbf{X};\mu,\sigma^2) = \prod_{i=1}^{N}\mathcal{N}(x_i|\mu,\sigma^2)$$

(b) What is the maximum likelihood (ML) estimation of the paramters, if both μ and σ^2 are unknown?
Solution:
Please refer to the lecture notes.

(c) Show that the ML estimation of the variance is biased, *i.e.* show that

$$\mathbb{E}(\sigma^2_{ML}) = \frac{N-1}{N}\sigma^2$$

Hint: you can use the fact that the expectation of random variable from a Gaussian distributuion is μ, *i.e.* $\mathbb{E}(x) = \mu$, and var$(x) = \sigma^2 = \mathbb{E}(x^2) - \mathbb{E}(x)^2$
Solution:

$$\begin{aligned}
\mathbb{E}(\sigma^2_{ML}) &= \mathbb{E}(\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_{ML})^2) \\
&= \mathbb{E}(\frac{1}{N}\left\{\sum_{i=1}^{N}(x_i^2 + \mu^2_{ML} - 2x_i\mu_{ML})\right\}) \\
&= \mathbb{E}(\frac{1}{N}\left\{\sum_{i=1}^{N}(x_i^2 + (\frac{1}{N}\sum_{j=1}^{N}x_j)^2 - 2x_i(\frac{1}{N}\sum_{j=1}^{N}x_j))\right\}) \\
&= \mathbb{E}(\frac{1}{N}\left\{\sum_{i=1}^{N}(x_i^2 + \frac{1}{N^2}\sum_{j=1}^{N}x_j^2 + \frac{1}{N^2}\sum_{j\neq k}x_jx_k - \frac{2}{N}x_i^2 - \frac{2}{N}\sum_{j=i}x_ix_j)\right\}) \\
&= \frac{1}{N}\left\{\sum_{i=1}^{N}\mathbb{E}(x_i^2) + \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathbb{E}(x_j^2) + \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j\neq k}\mathbb{E}(x_jx_k) - \frac{2}{N}\sum_{i=1}^{N}\mathbb{E}(x_i^2) - \frac{2}{N}\sum_{j\neq i}\mathbb{E}(x_ix_j)\right\} \\
&= \frac{1}{N}\left\{N(\sigma^2+\mu^2) + \sigma^2 + \mu^2 + (N-1)\mu^2 - 2(\sigma^2+\mu^2) - 2(N-1)\mu^2\right\} \\
&= \frac{1}{N}\left\{N\sigma^2 + N\mu^2 + \sigma^2 + \mu^2 + N\mu^2 - \mu^2 - 2\sigma^2 - 2\mu^2 - 2N\mu^2 + 2\mu^2\right\} \\
&= \frac{1}{N}(N\sigma^2 - \sigma^2) = \frac{N-1}{N}\sigma^2
\end{aligned}$$

(d) Please write down the objective function of maximum a posteriori (MAP) estimation of the parameters, if we assume that only μ is the unknown parameter and follow a Gaussian distribution with mean μ_0 and variance σ_0^2.
Solution:
We know that $p(\mu) \sim \mathcal{N}(\mu|\mu_0, \sigma_0^2)$, so the posterior probability of parameter μ would be:

$$\begin{aligned}
p(\mu|\mathbf{X}) &= p(\mathbf{X}|\mu)p(\mu) \\
&= \{\prod_{i=1}^{N}\mathcal{N}(x_i|\mu,\sigma^2)\}\mathcal{N}(\mu|\mu_0,\sigma_0^2)
\end{aligned}$$

(e) Please write down the Bayesian formulation of this problem, if all other assumptions stay the same as in question d.
Solution:

$$p(x'|\mathbf{X}) = \int p(x'|\theta)p(\theta|\mathbf{X})d\theta$$

where x' is some new data that you want to do prediction and θ is the set of parameters from the model, in this case $\theta = \{\mu\}$, since we assume μ is the only unknown parameter.