CSE 575 Homework 1

Andrew Dudley (addudley@asu.edu)

February 5, 2018

Question 1.

Suppose that in your coin flip experiment, you observed a set of α_H heads and α_T tails. Let θ denote the probability of observing heads, whose prior distribution follows $Beta(\beta_H, \beta_T)$, where β_H and β_T are two positive parameters.

(a) Prove that the posterior distribution $P(\theta|D)$ follows $Beta(\beta_H+\alpha_H, \beta_T+\alpha_T)$

The posterior distribution $P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$. We're told that

$$P(\theta) = \frac{\theta^{\beta_H-1}(1-\theta)^{\beta_T-1}}{B(\beta_H, \beta_T)} \sim Beta(\beta_H, \beta_T),$$

where $B(\beta_H, \beta_T)$ is the beta function.

Note that $P(D|\theta)$ is the likelihood function, which for a Bernoulli experiment is

$$\theta^{\alpha_H}(1-\theta)^{\alpha_T}.$$

Putting these together, we get

$$P(\theta|D) = \frac{\theta^{\beta_H+\alpha_H-1}(1-\theta)^{\beta_T+\alpha_T-1}}{P(D)B(\beta_H, \beta_T)},$$

and because both $P(D)$ and $B(\beta_H, \beta_T)$ are normalizing constants (they don't rely on θ), we can rewrite the equation as

$$P(\theta|D) = \frac{\theta^{\beta_H+\alpha_H-1}(1-\theta)^{\beta_T+\alpha_T-1}}{B(\beta_H+\alpha_H, \beta_T+\alpha_T)},$$

From here, we can see that

$$P(\theta|D) \sim Beta(\beta_H+\alpha_H, \beta_T+\alpha_T).$$

(b) What is the mean of $P(\theta|D)$?

For notational convenience, let $a = \beta_H + \alpha_H$ and $b = \beta_T + \alpha_T$. As they say, statistics is the practice of replacing expectations with averages, so the mean of $P(\theta|D)$ is the same as $\mathbb{E}\left[P(\theta|D)\right]$. And, of course, because this value is a probability, we can bound the integral by $[0, 1]$.

$$\begin{aligned}\mathbb{E}\left[P(\theta|D)\right] &= \int_0^1 \theta P(\theta|D) d\theta \\ &= \frac{1}{B(a,v)} \int_0^1 \theta^a (1-\theta)^{b-1} d\theta\end{aligned}$$

Here, we can see that the integral is, itself, a beta function.

$$\begin{aligned}&= \frac{B(a+1,b)}{B(a,b)} \\ &= \frac{\Gamma(a+1)\Gamma(b)\Gamma(a+b)}{\Gamma(a)\Gamma(b)\Gamma(a+b+1)} \\ &= \frac{a}{a+b} \\ &= \frac{\beta_H + \alpha_H}{\beta_H + \alpha_H + \beta_T + \alpha_T}\end{aligned}$$

(c) What is the MAP estimator $\hat{\theta}_{MAP}$ of θ?

$$\begin{aligned}\frac{d}{d\theta}\mathcal{L}(\theta) &= \frac{d}{d\theta} ln\left(\theta^{\beta_H + \alpha_H - 1}(1-\theta)^{\beta_T + \alpha_T - 1}\right) \\ &= (\beta_H + \alpha_H - 1)\left[\frac{d}{d\theta} ln\theta\right] + (\beta_T + \alpha_T - 1)\left[\frac{d}{d\theta} ln(1-\theta)\right] \\ &= \frac{\beta_H + \alpha_H - 1}{\theta} - \frac{\beta_T + \alpha_T - 1}{1-\theta}\end{aligned}$$

And to find a critical point (the minimum value), we set the derivate to 0.

$$\begin{aligned}\frac{\beta_H + \alpha_H - 1}{\theta} - \frac{\beta_T + \alpha_T - 1}{1-\theta} &= 0 \\ \hat{\theta}_{MAP} &= \frac{\beta_H + \alpha_H - 1}{\beta_H + \alpha_H + \beta_T + \alpha_T - 2}\end{aligned}$$

Question 2.

For this question, assume that $x_1, \cdots, x_N \in \mathbb{R}$ are i.i.d. from a normal distribution.

(a) Let $\hat{\mu}_{MLE}$ denote the MLE of μ. Prove that $\hat{\mu}_{MLE}$ is unbiased.

We'll first need to determine the equation for $\hat{\mu}_{MLE}$, then we'll need to compare its expectation to the population mean μ. Let $D = \{x_1, \cdots, x_n\}$. Then, because the data is assumed independent, the likelihood of μ with respect to D can be written as

$$P(D|\mu) = \Pi_{i=1}^{n} p(x_i|\mu)$$

Furthermore, we know that the samples are from a normal distribution, giving us

$$p(x|\mu) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} e^{\frac{(x-\mu)^2}{2\sigma^2}}$$

To simplify finding the derivative, take the log of the likelihood function

$$\mathcal{L}(P(D|\mu)) = \sum_{i=1}^{n} ln(p(x_i|\mu)$$

Now we'll find the MLE of μ by taking the derivative of the log-likelihood, setting it to 0, and solving for μ

$$\begin{aligned}
\frac{d\mathcal{L}}{d\mu} &= \sum_{i=1}^{n} \frac{d}{d\mu} ln(p(x_i|\mu)) && = 0 \\
&= -\sum_{i=1}^{n} \frac{d}{d\mu} \left(\frac{1}{2} ln(2\pi\sigma^2) + \frac{1}{2\sigma^2}(x_i - \mu)^2 \right) && = 0 \\
&= \sum_{i=1}^{n} (x_i - \hat{u}) && = 0 \\
\hat{\mu} &= \frac{1}{n} \sum_{i=1}^{n} x_i
\end{aligned}$$

Now that we know the equation for $\hat{\mu}$, we'll find its expectation.

$$\begin{aligned}\mathbb{E}[\hat{\mu}] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[x] \\ &= \frac{1}{n}\sum_{i=1}^{n}\mu \\ &= \mu\end{aligned}$$

Therefore, $\hat{\mu}_{MLE}$ is unbiased.

(b) If the true value of μ is unknown, then the MLE estimate of σ^2 is

$$\sigma^2_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu}_{MLE})^2$$

Prove that σ^2_{MLE} is biased.

Before we begin, there are a few derivations that we'll want to have on hand when completing this proof. To start, we'll derive the definition of variance σ^2.

$$\begin{aligned}\sigma^2 &= \mathbb{E}\left[(X-\mu)^2\right] \\ &= \mathbb{E}[X^2 - 2X\mu + \mu^2] \\ &= \mathbb{E}[X^2] - 2\mu\mathbb{E}[X] + \mu^2 \\ &= \mathbb{E}[X^2] - \mu^2 \\ &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2\end{aligned}$$

A little bit of algebraic manipulation of these equations gives us

$$\mathbb{E}[X^2] = \sigma^2 + \mu^2, \tag{1}$$

Also note that σ^2 is a function of X. If we pass into it $\hat{\mu}$, we get

$$\begin{aligned}\sigma^2_{\hat{\mu}} &= \mathbb{E}[\hat{\mu}^2] - (\mathbb{E}[\hat{\mu}])^2 \\ &= \mathbb{E}[\hat{\mu}^2] - \mu^2\end{aligned} \tag{2}$$

We can also derive the variance of $\hat{\mu}$ as follows

$$\begin{aligned}\sigma^2_{\hat{\mu}} &= Var(\frac{x_1 + \cdots + x_n}{n}) \\ &= \frac{1}{n^2} Var(x_1 + \cdots + x_n) \\ &= \frac{1}{n^2}[Var(x_1) + \cdots + Var(x_n)] \\ &= \frac{\sigma^2}{n} \end{aligned} \tag{3}$$

Using (2) and (3), we arrive at the equation

$$\mathbb{E}\left[\hat{\mu}^2\right] = \frac{\sigma^2}{n} + \mu^2 \tag{4}$$

With these derivations in our toolbox, we can now begin the proof that σ^2_{MLE} is biased. Again, we'll do so by comparing the expectation of this value to the population variance.

$$\begin{aligned}\mathbb{E}\left[\sigma^2_{MLE}\right] &= \mathbb{E}\left[\frac{1}{n}\sum (x_i - \hat{\mu})^2\right] \\ &= \frac{1}{n}\mathbb{E}\left[\sum \left(x_i^2 - 2x_i\hat{\mu} + \hat{\mu}^2\right)\right] \\ &= \frac{1}{n}\mathbb{E}\left[\sum x_i^2 - 2\hat{\mu}\sum x_i + N\hat{\mu}^2\right] \\ &= \frac{1}{n}\end{aligned}$$

Question 3.

(a) How many independent parameters would there be for the Naive Bayes classifier trained with the given data? What are they?

There would be **thirteen independent parameters**. Note first that RID is a nominal attribute and buys_computer is the dependent (target) attribute, leaving us with age, income, student, and credit_rating as the attributes of interest.

An assumption is made with Naive Bayes that

$$P(X_i|X_{1,\ldots,i-1,x_i+1,\ldots,n}Y) = P(X_i|Y) \forall\, i \in \{1, \cdots, n\}$$

A Bayes classifier can be represented as

$$P(Y_c|X_{1..n}) = \frac{P(X_{1..n}|Y_c)P(Y_c)}{\sum_Y P(X_{1..n}|Y)P(Y)}$$

Given the conditional independence assumption of Naive Bayes, this can be simplified to

$$P(Y_c|X_{1..n}) = \frac{\Pi_i P(X_i|Y_c)P(Y_c)}{\sum_Y P(X_i|Y)P(Y)}$$

Let m_i represent the number of discrete categories in the variable x_i and C represent the number of discrete classes of the target attribute Y. Then, given $P(X = X_{ij}|Y = y_c)$ where $j \in \{1, \cdots, m_i\}$ and $c \in \{1, \cdots, C\}$, each independent variable x_i contributes $m_i - 1$ independent parameters for each class y_c. Note that, given $m_i - 1$ parameters for the class conditional of an attribute, the final parameter of that class conditional is simply the difference between 1 and the sum of those parameters, and is therefore not independent.

Similarly, we must account for the independent parameters contributed by the prior $P(Y)$, which will be $C - 1$ parameters.

Thus, the equation for the total number of independent parameters in a Naive Bayes model will be

$$C * \sum_{i=1}^{n}(m_i - 1) + C - 1$$

Plugging in the attributes from the data provided, the number of independent parameters is **thirteen**.

(b)

$$P(age = "youth"|buys\_computer = "no") = 3/5, P(age = "middle\_aged"|buys\_computer = "no") = 0$$
$$P(age = "youth"|buys\_computer = "yes") = 2/9, P(age = "middle\_aged"|buys\_computer = "yes") = 4/9$$
$$P(income = "low"|buys\_computer = "no") = 1/5, P(income = "medium"|buys\_computer = "no") = 2/5$$
$$P(income = "low"|buys\_computer = "yes") = 3/9, P(income = "medium"|buys\_computer = "yes") = 4/9$$
$$P(student = "no"|buys\_computer = "no") = 4/5$$
$$P(student = "no"|buys\_computer = "yes") = 3/9$$
$$P(credit\_rating) = "fair"|buys\_computer = "no") = 2/5$$
$$P(credit\_rating) = "fair"|buys\_computer = "yes") = 6/9$$
$$P(buys\_computer = "yes") = 9/14$$

(c) Given a new person with features $x = (youth, medium, yes, fair)$, what is $P(Y = yes|x)$?

$$\frac{\frac{2}{9} * \frac{4}{9} * \frac{6}{9} * \frac{6}{9} * \frac{9}{14}}{\frac{2}{9} * \frac{4}{9} * \frac{6}{9} * \frac{6}{9} * \frac{9}{14} + \frac{3}{5} * \frac{2}{5} * \frac{1}{5} * \frac{2}{5} * \frac{5}{14}} = 0.80451$$

Therefore, NB would classify this input at $Y = yes$.

Question 4.

Suppose we have two positive examples $x_1 = (1, 0)$ and $x_2 = (0, -1)$, and two negative samples $x_3 = (0, 1)$ and $x_4 = (-1, 0)$. Apply standard gradient ascent method to train a logistic regression classifier (without any regularization term). Initialize the weight vector with two different value and set $w_0^0 = 0$. Would the final weight vector (w^*) be the same for the two different initial values? What are the values? You may assume the learning rate to be a positive real constant η.

For notational and implementation convenience, first append a 1 to each sample vector

$$\left\{ X = \begin{bmatrix} 1 & 0 & 1 \\ 0 & -1 & 1 \\ 0 & 1 & 1 \\ -1 & 0 & 1 \end{bmatrix} \right\}_{\text{m}}, \quad \theta = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

$$\mathbf{z} = X\theta = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \hat{\mathbf{p}} = sig((z)) = \frac{1}{1 + e^{-z}} = \begin{bmatrix} 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \end{bmatrix}$$

$\hat{\mathbf{p}}$ represents the model's estimated probabilities of a positive class in vectorized form. When performing prediction, a threshold value of 0.5 is set such that the predicted class is 1 when $p > 0.5$ and 0 when $p \leq 0.5$. This threshold effectively draws a discriminant line in the input space. When training, however, The cross-entropy loss is used as a cost function on the estimated probability. When training with batches of data, the total loss is simply the average cost over the training data. This allows us to write the cost function as

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} y^{(i)} log(\hat{p}^{(i)}) + (1 - y^{(i)}) log(1 - \hat{p}^{(i)})$$

To minimize the loss, we can calculate the gradient of $J(\theta)$ with respect to θ and then update θ in the negative direction of that gradient. Iterating over

this process ad infinitum will converge to the optimal parameters (ignoring issues of perfect separation, etc).

$$\frac{\delta J(\theta)}{\delta \theta_j} = \frac{\delta J}{\delta \hat{p}} \frac{\delta \hat{p}}{\delta z} \frac{\delta z}{\theta_j}$$

$$\begin{aligned} \frac{\delta J}{\delta \hat{p}} &= -\frac{1}{m} \sum_{i=1}^{m} \left[\frac{y^{(i)}}{\hat{p}^{(i)}} - \frac{1 - y^{(i)}}{1 - \hat{p}^{(i)}} \right] \\ &= -\frac{1}{m} \sum_{i=1}^{m} \left[\frac{y^{(i)} - \hat{p^{(i)}}}{\hat{p}^{(i)}(1 - \hat{p}^{(i)})} \right] \end{aligned}$$

$$\begin{aligned} \frac{\delta \hat{p}}{\delta z} &= sig(z)(1 - sig(z)) \\ &= \hat{p}(1 - \hat{p}) \end{aligned}$$

$$\frac{\delta J}{\delta \hat{p}} \frac{\delta \hat{p}}{\delta z} = \frac{1}{m} \sum_{i=1}^{m} \hat{p}^{(i)} - y^{(i)}$$

$$\frac{\delta z}{\delta \theta_j} = x_j$$

$$\frac{\delta J}{\delta \hat{p}} \frac{\delta \hat{p}}{\delta z} \frac{\delta z}{\theta_j} = \frac{1}{m} \sum_{i=1}^{m} (\hat{p}^{(i)} - y^{(i)}) x_j^{(i)}$$

From this result, it is clear that the only calculation we need to complete prior to updating our weights is the sigmoid function. It's also worth noting that this result can be written in matrix form, which allows for efficient computation when implementing

$$\frac{\delta J}{\delta \hat{p}} \frac{\delta \hat{p}}{\delta z} \frac{\delta z}{\theta} = mean((sig(X\theta) - \mathbf{y}) \odot X)$$

Now that we know the process of logistic regression, what can we say about the parameters? What do they represent?

Unfortunately, interpreting the parameters learned isn't nearly as straight forward as it is in linear regression models, even though what we're ultimately fitting is a discriminant line in the feature space. However, if we consider the threshold value 0.5 and its relation of the the input to the sigmoid function z, we know that when $z > 0$, $sig(z) > 0.5$. Likewise, when $z \leq 0$, $sig(z) \leq 0.5$.

From this, it is obvious that the parameters we are learning are for a line in the feature space at form at $z = 0$. Scaling the coefficients of z by any scalar value would result in the same line, so the there are infinite values that θ could converge to, and this will depend on the the values that the vector is initiated with.

In this example, the values will ultimately converge to $\theta_1 = -\theta_2$ (s.t. $\theta_1 \in \mathcal{R}$), and $b = 0$

$$\theta_1 x_1 + \theta_2 x_2 + b = 0$$