CS391L HW3: Problem Set

Mingyo Seo,
UT EID: ms84662
Email: mingyo@utexas.edu

I. BASIC PROBABILITY

The **answer** is $p(X|Y) = \frac{1}{73}$. For the notation, please see the below description.

Let X be the case that it will actually rain today, and Y be the case that the meterologist predicts a rainy day. In other words, X^c means that it will not rain today. In the case of Y^c, the meterologist predicts that it would not rain today. From the given historical data,

$$p(X) = \frac{5}{365} \quad = \frac{1}{73}, \tag{1}$$

$$p(X^c) = 1 - P(X) \quad = \frac{72}{73}. \tag{2}$$

The meterologist correctly predicts 90% when it rains, so

$$p(Y|X) = \frac{9}{10}, \tag{3}$$

The meterologist correctly predicts 10% when it doesn't rain, so

$$p(Y^c|X^c) = \frac{1}{10}, \tag{4}$$

This also implies that

$$p(Y|X^c) = 1 - p(Y^c|X^c) \quad = \frac{9}{10}. \tag{5}$$

Given the condition that a meterologist predicts rain today, the conditional probability of rain today can be formulated as,

$$p(X|Y) = \frac{p(X,Y)}{p(Y)} \quad = \frac{p(X,Y)}{p(X,Y) + p(X^c,Y)}. \tag{6}$$

From Bayes' theorem, we can get

$$p(X,Y) = p(Y|X)p(X), \tag{7}$$

$$p(X^c,Y) = p(Y|X^c)p(X^c). \tag{8}$$

Therefore, from Equation 1, 2, 3, 5, we can compute the probabilities,

$$p(X,Y) = \frac{9}{10} \times \frac{1}{73}, \tag{9}$$

$$p(X^c,Y) = \frac{9}{10} \times \frac{72}{73}. \tag{10}$$

Therefore, we can compute the **answer** by plugging these probabilities in 6, as

$$p(X|Y) = \frac{p(X,Y)}{p(X,Y) + p(X^c,Y)} \quad = \frac{\frac{9}{10} \times \frac{1}{73}}{\left(\frac{9}{10} \times \frac{1}{73}\right) + \left(\frac{9}{10} \times \frac{72}{73}\right)} = \frac{1}{73}. \tag{11}$$

II. Entropy

From the definitions,

$$H[\boldsymbol{y}] = -\int p(\boldsymbol{y}) \ln p(\boldsymbol{y}) dy, \tag{12}$$

$$H[\boldsymbol{y}|\boldsymbol{x}] = -\int\int p(\boldsymbol{x},\boldsymbol{y}) \ln p(\boldsymbol{y}|\boldsymbol{x}) d\boldsymbol{y} d\boldsymbol{x}. \tag{13}$$

From Bayes' theorem, we can get

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{p(\boldsymbol{x},\boldsymbol{y})}{p(\boldsymbol{x})}, \tag{14}$$

$$p(\boldsymbol{x},\boldsymbol{y}) = p(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{y}). \tag{15}$$

Therefore, we can re-write the given KL definition of mutual informaiton, as

$$\begin{aligned} I[\boldsymbol{x},\boldsymbol{y}] &= -\int\int p(\boldsymbol{x},\boldsymbol{y}) \ln \left\{ \frac{p(\boldsymbol{x})p(\boldsymbol{y})}{p(\boldsymbol{x},\boldsymbol{y})} \right\} d\boldsymbol{x} d\boldsymbol{y} \\ &= -\int\int p(\boldsymbol{x},\boldsymbol{y}) \ln \frac{p(\boldsymbol{y})}{p(\boldsymbol{y}|\boldsymbol{x})} d\boldsymbol{x} d\boldsymbol{y} \\ &= -\int\int p(\boldsymbol{x},\boldsymbol{y}) \left\{ \ln p(\boldsymbol{y}) - \ln p(\boldsymbol{y}|\boldsymbol{x}) \right\} d\boldsymbol{x} d\boldsymbol{y} \\ &= -\int\int p(\boldsymbol{x},\boldsymbol{y}) \ln p(\boldsymbol{y}) d\boldsymbol{x} d\boldsymbol{y} + \int\int p(\boldsymbol{x},\boldsymbol{y}) \ln p(\boldsymbol{y}|\boldsymbol{x}) d\boldsymbol{x} d\boldsymbol{y}, \\ &= -\int\int p(\boldsymbol{y})p(\boldsymbol{x}|\boldsymbol{y}) \ln p(\boldsymbol{y}) d\boldsymbol{x} d\boldsymbol{y} + \int\int p(\boldsymbol{x},\boldsymbol{y}) \ln p(\boldsymbol{y}|\boldsymbol{x}) d\boldsymbol{x} d\boldsymbol{y}. \end{aligned} \tag{16}$$

By using the property,

$$\forall \boldsymbol{y}, \int p(\boldsymbol{x}|\boldsymbol{y}) d\boldsymbol{x} = 1, \tag{17}$$

we can re-write the first term of the above equation with respect to $H[\boldsymbol{x}]$ at Equation 12, as

$$\begin{aligned} -\int\int p(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{y}) \ln p(\boldsymbol{y}) d\boldsymbol{x} d\boldsymbol{y} &= -\int p(\boldsymbol{y}) \ln p(\boldsymbol{y}) \left\{ \int p(\boldsymbol{x}|\boldsymbol{y}) d\boldsymbol{x} \right\} d\boldsymbol{y} \\ &= -\int p(\boldsymbol{y}) \ln p(\boldsymbol{y}) d\boldsymbol{y} \\ &= H[\boldsymbol{y}]. \end{aligned} \tag{18}$$

The second term of Equation 16 can be re-written with respect to $H[\boldsymbol{y}|\boldsymbol{x}]$ at Equation 13 as

$$\begin{aligned} \int\int p(\boldsymbol{x},\boldsymbol{y}) \ln p(\boldsymbol{y}|\boldsymbol{x}) d\boldsymbol{x} d\boldsymbol{y} &= -\left(-\int\int p(\boldsymbol{x},\boldsymbol{y}) \ln p(\boldsymbol{y}|\boldsymbol{x}) d\boldsymbol{x} d\boldsymbol{y} \right) \\ &= -\left(-\int\int p(\boldsymbol{x},\boldsymbol{y}) \ln p(\boldsymbol{y}|\boldsymbol{x}) d\boldsymbol{y} d\boldsymbol{x} \right) \\ &= -H[\boldsymbol{y}|\boldsymbol{x}]. \end{aligned} \tag{19}$$

Therefore, we get the results,

$$I[\boldsymbol{x},\boldsymbol{y}] = H[\boldsymbol{y}] - H[\boldsymbol{y}|\boldsymbol{x}]. \tag{20}$$

Due to the symmetric form of $I[\boldsymbol{x},\boldsymbol{y}]$, by switching $\boldsymbol{x}$ and $\boldsymbol{y}$ at Equation 16, we can also get

$$\begin{aligned} I[\boldsymbol{x},\boldsymbol{y}] &= I[\boldsymbol{y},\boldsymbol{x}] \\ &= H[\boldsymbol{x}] - H[\boldsymbol{x}|\boldsymbol{y}]. \end{aligned} \tag{21}$$

III. Beta Distribution

From the definitions Beta distribution,

$$p(\mu) = \frac{\mu^{a-1}\left(1-\mu\right)^{b-1}}{C(a,b)}, \tag{22}$$

where C(a,b) is the normalizing constant, given as

$$\begin{aligned} C(a,b) &= \int_0^1 \mu^{a-1}\left(1-\mu\right)^{b-1} d\mu, \\ &= \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}. \end{aligned} \tag{23}$$

Then, $E(\mu)$ is given as,

$$\begin{aligned} E(\mu) &= \int_0^1 \mu p(\mu) d\mu \\ &= \int_0^1 \mu \frac{\mu^{a-1}\left(1-\mu\right)^{b-1}}{C(a,b)} d\mu \\ &= \frac{1}{C(a,b)} \int_0^1 \mu^{a}\left(1-\mu\right)^{b-1} d\mu. \end{aligned} \tag{24}$$

The integration term of Equation 24, $\int_0^1 \mu^{a}\left(1-\mu\right)^{b-1} d\mu$, also follows the cumulative distribution formulation of the Beta distribution. Therefore, we can further re-write Equation 24 term as,

$$\begin{aligned} \int_0^1 \mu^{a}\left(1-\mu\right)^{b-1} d\mu &= \int_0^1 \mu^{(a+1)-1}\left(1-\mu\right)^{b-1} d\mu \\ &= \int_0^1 \mu^{(a+1)-1}\left(1-\mu\right)^{b-1} d\mu \\ &= C(a+1,b) \\ &= \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)}. \end{aligned} \tag{25}$$

By plugging the above term in Equation 24, we can re-write $E(\mu)$, as

$$\begin{aligned} E(\mu) &= \frac{1}{C(a,b)} \int_0^1 \mu^{a}\left(1-\mu\right)^{b-1} d\mu \\ &= \frac{1}{C(a,b)} \times \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \times \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \times \frac{a\Gamma(a)\Gamma(b)}{(a+b)\Gamma(a+b)} \\ &= \frac{a}{a+b}. \end{aligned} \tag{26}$$

IV. SUPPORT VECTOR MACHINES

Recall that the goal of a support vector machine (SVM) classifier is maximizing the geometric margin ρ, given data $\boldsymbol{x}_i$ with labels y_i,

$$\max_{\boldsymbol{w},b} \rho = \min_i \frac{y_i(\boldsymbol{w}^\top \phi(\boldsymbol{x}_i) + b)}{\|\boldsymbol{w}\|}. \tag{27}$$

Here, $i \in \{1, 2, ..., N\}$, N is the number of data, and $\phi(\cdot)$ is a kernel function that maps $\boldsymbol{x}_i$ into a hyperspace of more dimensions, and b is the bias. By scaling as $\boldsymbol{w} \to \kappa \boldsymbol{w}$ and $b \to \kappa b$, we can get

$$\min_i y_i(\boldsymbol{w}^\top \phi(\boldsymbol{x}_i) + b) = 1, \tag{28}$$

and re-formulate the problem as,

$$\max_{\boldsymbol{w},b} \rho = \frac{1}{\|\boldsymbol{w}\|} \tag{29}$$

$$\text{s.t. } y_i(\boldsymbol{w}^\top \phi(\boldsymbol{x}_i) + b) \leq 1. \tag{30}$$

Then, this optimization can be transformed as

$$\max_{\boldsymbol{w},b} \rho \Leftrightarrow \min_{\boldsymbol{w},b} \frac{1}{\rho^2} \Leftrightarrow \max_{\boldsymbol{w},b} \frac{1}{2} \|\boldsymbol{w}\|^2 \tag{31}$$

where the same constraint of Equation 30 is given.

By using the method of Lagrange multipliers, Equation 30, 31 are formulated, as

$$\min_{\boldsymbol{w},b,\boldsymbol{\lambda}} L(\boldsymbol{w}, b, \boldsymbol{\lambda}) \tag{32}$$

where

$$L(\boldsymbol{w}, b, \boldsymbol{\lambda}) = \frac{1}{2} \|\boldsymbol{w}\|^2 + \sum_{i=1}^{N} \lambda_i \left(y_i(\boldsymbol{w}^\top \phi(\boldsymbol{x}_i) + b) - 1 \right). \tag{33}$$

The solutions of the problem are computed from the partial derivatives,

$$\frac{\partial L}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{i=1}^{N} \lambda_i y_i \phi(\boldsymbol{x}_i) = 0, \tag{34}$$

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{N} \lambda_i y_i = 0. \tag{35}$$

Also, a constrained optimization must satisfy the Karush-Kuhn-Tucker (KKT) conditions [1], which yields the conditions,

$$\begin{aligned} \forall i, \quad & \lambda_i > 0, \\ & \lambda_i \left(y_i(\boldsymbol{w}^\top \phi(\boldsymbol{x}_i) + b) - 1 \right) = 0. \end{aligned} \tag{36}$$

By using Equation 36, we can get the following equation,

$$\sum_{i=1}^{N} \lambda_i \left(y_i(\boldsymbol{w}^\top \phi(\boldsymbol{x}_i) + b) - 1 \right) = 0. \tag{37}$$

This yields

$$\begin{aligned} 0 &= \sum_{i=1}^{N} \lambda_i y_i \boldsymbol{w}^\top \phi(\boldsymbol{x}_i) + \sum_{i=1}^{N} \lambda_i y_i b - \sum_{i=1}^{N} \lambda_i \\ &= \boldsymbol{w}^\top \sum_{i=1}^{N} \lambda_i y_i \phi(\boldsymbol{x}_i) + b \sum_{i=1}^{N} \lambda_i y_i - \sum_{i=1}^{N} \lambda_i. \end{aligned} \tag{38}$$

By plugging Equation 34, 35 in the above equation, we can get

$$\begin{aligned} 0 &= \boldsymbol{w}^\top \sum_{i=1}^{N} \lambda_i y_i \phi(\boldsymbol{x}_i) + b \sum_{i=1}^{N} \lambda_i y_i - \sum_{i=1}^{N} \quad = \boldsymbol{w}^\top \boldsymbol{w} + b \times 0 - \sum_{i=1}^{N} \lambda_i \\ &= \|\boldsymbol{w}\|^2 - \sum_{i=1}^{N} \lambda_i. \end{aligned} \tag{39}$$

Thus, we can show that

$$\|\boldsymbol{w}\|^2 = \sum_{i=1}^{N} \lambda_i. \tag{40}$$

V. Gibbs Sampling

The Gaussian random variable ν is independent from z_i, and $E(\nu) = 0$. Therefore, we can get

$$E(\sigma_i \nu) = E(\sigma_i)E(\nu) = E(\sigma_i) \times 0 = 0. \tag{41}$$

Therefore, the expectation of $\acute{z}_i$ can be re-written, as

$$\begin{aligned} E(\acute{z}_i) &= E\left(\mu_i + \alpha(z_i - \mu_i) + \sigma_i(1-\alpha^2)^{\frac{1}{2}}\nu\right) \\ &= (1-\alpha)E(\mu_i) + \alpha E(z_i) + (1-\alpha^2)^{\frac{1}{2}} E(\sigma_i \nu) \\ &= (1-\alpha)E(\mu_i) + \alpha E(z_i) + (1-\alpha^2)^{\frac{1}{2}} \times 0 \\ &= (1-\alpha)E(\mu_i) + \alpha E(z_i). \end{aligned} \tag{42}$$

From the definition, $\mu_i = E(z_i)$. Also, $E(\mu_i) = \mu_i$. Thus, we can further simplify Equation 42, as

$$\begin{aligned} E(\acute{z}_i) &= (1-\alpha)\mu_i + \alpha E(z_i) \\ &= (1-\alpha)\mu_i + \alpha \mu_i \\ &= \mu_i. \end{aligned} \tag{43}$$

Therefore, the mean of $\acute{z}_i$ is also μ_i.

References

[1] Christopher M. Bishop (2006) *Pattern Recognition and Machine Learning*, Springer.