Home Assignment №1 Solutions

October 20, 2024

Exercise 1

[5 points]. This problem reviews basic concepts from probability.

a) [1 point]. A biased die has the following probabilities of landing on each face:

face	1	2	3	4	5	6
P(face)	.1	.1	.2	.2	.4	0

I win if the die shows even. What is the probability that I win? Is this better or worse than a fair die (i.e., a die with equal probabilities for each face)?

Solution:
$P[\text{even}] = P(2) + P(4) + P(6) = 0.1 + 0.2 + 0 = 0.3$. This is *worse* than a fair die which has probability 0.5 to land on an even number.

b) [1 point]. Recall that the expected value $\mathbb{E}[X]$ for a random variable X is

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} p(X = x)\ x,$$

where $\mathcal{X}$ is the set of values X may take on. Similarly, the expected value of any function f of random variable X is

$$\mathbb{E}[f(X)] = \sum_{x \in \mathcal{X}} p(X = x)\ f(x).$$

Now consider the function below, which we call the "indicator function"

$$\mathbb{I}[X = a] := \begin{cases} 1 & \text{if } X = a \\ 0 & \text{if } X \neq a \end{cases}.$$

Let X be a random variable which takes on the values $3, 8$ or 9 with probabilities p_3, p_8 and p_9 respectively. Calculate $\mathbb{E}[\mathbb{I}[X = 8]]$.

Solution:

$$\mathbb{E}[\mathbb{I}[X = 8]] = \sum_{x \in \{3,8,9\}} p_x \mathbb{I}[X = 8] = p_3 \times 0 + p_8 \times 1 + p_9 \times 0 = p_8.$$

c) [2 points]. Recall the following definitions:

- Entropy: $H(X) = -\sum_{x \in \mathcal{X}} p(X = x) \log_2 p(X = x) \quad = -\mathbb{E}[\log_2 p(X)]$
- Joint entropy: $H(X, Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(X = x, Y = y) \log_2 p(X = x, Y = y) \quad = -\mathbb{E}[\log_2 p(X, Y)]$
- Conditional entropy: $H(Y|X) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(X = x, Y = y) \log_2 p(Y = y|X = x) \quad = -\mathbb{E}[\log_2 p(Y|X)]$
- Mutual information: $I(X; Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(X = x, Y = y) \log_2 \frac{p(X=x,Y=y)}{p(X=x)p(Y=y)}$

Using the definitions of the entropy, joint entropy, and conditional entropy, prove the following chain rule for the entropy:

$$H(X, Y) = H(Y) + H(X|Y).$$

Solution:

$$\begin{aligned}
H(X, Y) &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(X = x, Y = y) \log_2 p(X = x, Y = y) \\
&= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(X = x, Y = y) \log_2 p(X = x) p(Y = y|X = x) \\
&= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(X = x, Y = y) \log_2 p(X = x) \\
&\quad -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(X = x, Y = y) \log_2 p(Y = y|X = x) \\
&= -\sum_{x \in \mathcal{X}} p(X = x) \log_2 p(X = x) \\
&\quad -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(X = x, Y = y) \log_2 p(Y = y|X = x) \\
&= H(X) + H(Y|X).
\end{aligned}$$

d) [1 point]. Recall that two random variables X and Y are *independent* if

$$\text{for all } x \in \mathcal{X} \text{ and all } y \in \mathcal{Y}, \quad p(X = x, Y = y) \;=\; p(X = x)p(Y = y).$$

If variables X and Y are independent, is $I(X;Y) = 0$? If yes, prove it. If no, give a counter example.

Solution:

Since variables X and Y are independent

$$\begin{aligned} I(X;Y) &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(X=x,Y=y)\log_2 \frac{p(X=x,Y=y)}{p(X=x)p(Y=y)} \\ &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(X=x,Y=y)\log_2 \frac{p(X=x)p(Y=y)}{p(X=x)p(Y=y)} \\ &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(X=x,Y=y)\log_2 1 \\ &= 0. \end{aligned}$$

Exercise 2

[4 points]. Given a training set $\mathcal{D} = \{(x^{(i)}, y^{(i)}), i = 1, \ldots, M\}$, where $x^{(i)} \in \mathbb{R}^N$ and $y^{(i)} \in \{1, 2, \ldots, C\}$, derive the maximum likelihood estimates of the naive Bayes for real valued $x_j^{(i)}$ modeled with a Laplacian distribution, *i.e.*,

$$p(x_j|y = c) = \frac{1}{2\sigma_{j|c}} \exp\left(-\frac{|x_j - \mu_{j|c}|}{\sigma_{j|c}}\right).$$

Solution:

Proof. Given a training set $\mathcal{D} = \{(x^{(i)}, y^{(i)}), i = 1, \cdots, M\}$, we write down the joint probability distribution of the data

$$\begin{aligned} p(\mathcal{D};\phi,\theta) &= \prod_{i=1}^{M} p(x^{(i)}, y^{(i)};\phi,\theta) \\ &= \prod_{i=1}^{M} p(y^{(i)};\phi)p(x^{(i)}|y^{(i)};\theta) \\ &= \prod_{i=1}^{M} p(y^{(i)};\phi)\prod_{j=1}^{N} p(x_j^{(i)}|y^{(i)};\theta_{j|c}). \end{aligned} \tag{1}$$

When we wish to explicitly view this as a function of the parameters ϕ and θ, we instead call it the likelihood function of the data $L(\phi, \theta)$. The principal of maximum likelihood says that we should choose ϕ, θ so as to make the data as high probability as possible. That is, we should choose ϕ, θ to maximize $L(\phi, \theta)$. Instead of maximizing $L(\phi, \theta)$, we can also maximize any strictly increasing function of $L(\phi, \theta)$. In particular, the derivations will be a bit simpler if we instead maximize the log likelihood

$$\begin{aligned}
\ell(\phi, \theta) &= \sum_{i=1}^{M} \log p(y^{(i)}; \phi) + \sum_{i=1}^{M} \sum_{j=1}^{N} \log p(x_j^{(i)} | y^{(i)}; \theta_{j|c}) \\
&= \sum_{i=1}^{M} \sum_{y^{(i)} \in \{1,2,\ldots,C\}} \mathbb{I}[y^{(i)} = c] \log \phi_y + \sum_{i=1}^{M} \sum_{j=1}^{N} \log p(x_j^{(i)} | y^{(i)}; \theta_{j|c}), \quad (2)
\end{aligned}$$

For real valued x_j, we model it with a Laplacian distribution

$$p(x_j | y = c) = \frac{1}{2\sigma_{j|c}} \exp\left(-\frac{|x_j - \mu_{j|c}|}{\sigma_{j|c}}\right).$$

If we pick out all terms in Eq. (2) that depend only on $\mu_{j|c}$, $\sigma_{j|c}$, we have

$$J(\mu_{j|c}, \sigma_{j|c}) = \sum_{i=1}^{M} \mathbb{I}[y^{(i)} = c] \left(-\log 2\sigma_{j|c} - \frac{|x_j - \mu_{j|c}|}{\sigma_{j|c}}\right). \quad (3)$$

Since it is the extreme problem of the location parameter for Laplace distribution, when $\mu_{j|c}$ is the median, the derivative w.r.t. $\mu_{j|c}$ will be zero.

Taking the derivative w.r.t. $\sigma_{j|c}$ and setting it to zero, we have

$$\begin{aligned}
&\sum_{i=1}^{M} \mathbb{I}[y^{(i)} = c] \left(-\frac{1}{\sigma_{j|c}} + \frac{|x_j^{(i)} - \mu_{j|c}|}{\sigma_{j|c}^2}\right) = 0 \\
&\sum_{i=1}^{M} \mathbb{I}[y^{(i)} = c] \left(-1 + \frac{|x_j^{(i)} - \mu_{j|c}|}{\sigma_{j|c}}\right) = 0 \\
&\sigma_{j|y} = \frac{\sum_{i=1}^{M} \mathbb{I}[y^{(i)} = c] |x_j^{(i)} - \mu_{j|c}|}{\sum_{i=1}^{M} \mathbb{I}[y^{(i)} = c]}. \quad (4)
\end{aligned}$$

□

Exercise 3

[4 points]. Prove that in binary classification, the posterior of linear discriminant analysis, *i.e.*, $p(y = 1|x; \phi, \mu, \Sigma)$, admits a sigmoid form

$$p(y = 1|x; \theta) = \frac{1}{1 + e^{-\theta^T x}}, \tag{5}$$

where θ is a function of $\{\phi, \mu, \Sigma\}$. Hint: remember to use the convention of letting $x_0 = 1$.

Solution:

Proof. Making use of the Bayes' rule, the law of total probability, and the chain rule of probability, we have

$$\begin{aligned} p(y = 1|x) &= \frac{p(x, y = 1)}{p(x)} & (6)\\ &= \frac{p(x|y = 1)p(y = 1)}{p(x|y = 0)p(y = 0) + p(x|y = 1)p(y = 1)} & (7)\\ &= \frac{1}{1 + \frac{p(x|y=0)p(y=0)}{p(x|y=1)p(y=1)}}. & (8) \end{aligned}$$

This equation seems very much like what we are looking for. Let's take a closer look at the fraction

$$\begin{aligned} \frac{p(x|y = 0)p(y = 0)}{p(x|y = 1)p(y = 1)} &= \frac{(1 - \phi) \exp\left\{-\frac{1}{2}(x - \mu_0)^T \Sigma^{-1}(x - \mu_0)\right\}}{\phi \exp\left\{-\frac{1}{2}(x - \mu_1)^T \Sigma^{-1}(x - \mu_1)\right\}} \\ &= \exp\left[\log \frac{1 - \phi}{\phi} - \frac{1}{2}(x - \mu_0)^T \Sigma^{-1}(x - \mu_0) + \frac{1}{2}(x - \mu_1)^T \Sigma^{-1}(x - \mu_1)\right] \\ &= \exp\left[\left(\log \frac{1 - \phi}{\phi} - \frac{1}{2}\mu_0^T \Sigma^{-1}\mu_0 + \frac{1}{2}\mu_1^T \Sigma^{-1}\mu_1\right) x_0 + (\mu_0 - \mu_1)^T \Sigma^{-1} x\right], \end{aligned} \tag{9}$$

where we let $x_0 = 1$. Therefore, we have

$$p(y = 1|x; \theta) = \frac{1}{1 + e^{-\theta^T x}}, \tag{10}$$

where

$$\theta = \begin{bmatrix} -\left(\log \frac{1-\phi}{\phi} - \frac{1}{2}\mu_0^T \Sigma^{-1}\mu_0 + \frac{1}{2}\mu_1^T \Sigma^{-1}\mu_1\right) \\ -\Sigma^{-1}(\mu_0 - \mu_1) \end{bmatrix}. \tag{11}$$

□

Exercise 4

[2 points]. For an N-dimensional vector x, the multivariate Gaussian distribution takes the form

$$\mathcal{N}(x;\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^N|\Sigma|}} \exp\left\{-\frac{1}{2}\left(x-\mu\right)^T \Sigma^{-1}\left(x-\mu\right)\right\}. \tag{12}$$

We partition x into two disjoint subsets x_a and x_b. Without loss of generality, we can take x_a to form the first N_1 elements of x, with x_b comprising the remaining $N - N_1$ elements such that

$$x = \begin{bmatrix} x_a \\ x_b \end{bmatrix}, \tag{13}$$

$$\mu = \begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix}, \tag{14}$$

and

$$\Sigma^{-1} = \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix}^{-1} = \begin{bmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{bmatrix}, \tag{15}$$

where $\Sigma_{ab}^T = \Sigma_{ba}$ and $\Lambda_{ab}^T = \Lambda_{ba}$. Prove that the conditional of a joint Gaussian distribution $x_b|x_a$ given by

$$p\left(x_b|x_a\right) = \frac{p\left(x_a, x_b;\mu,\Sigma\right)}{\int p\left(x_a, x_b;\mu,\Sigma\right) dx_b} \tag{16}$$

is also Gaussian.

<u>Hints:</u> You may derive the mean vector and the covariance matrix of $p(x_b|x_a)$ by comparing the coefficients of your expression with the following general form:

$$\frac{1}{2}z^T A z + b^T z + c = \frac{1}{2}\left(z + A^{-1}b\right)^T A\left(z + A^{-1}b\right) + c - \frac{1}{2}b^T A^{-1} b. \tag{17}$$

By the way, the method is called "completing the square".

Besides, you may find this more general result of block matrix inverse relating to Eq. (15) useful for interpreting your solution:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} M & -MBD^{-1} \\ -D^{-1}CM & D^{-1} + D^{-1}CMBD^{-1} \end{bmatrix} \tag{18}$$

where we have defined

$$M = (A - BD^{-1}C)^{-1}. \tag{19}$$

Solution:

Proof.

$$p\left(x_b|x_a\right) = \frac{p\left(x_a, x_b; \mu, \Sigma\right)}{\int p\left(x_a, x_b; \mu, \Sigma\right) dx_b} \tag{20}$$

$$= \frac{1}{Z'} \exp\left(-\frac{1}{2}\begin{bmatrix} x_a - \mu_a \\ x_b - \mu_b \end{bmatrix}^T \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix}^{-1} \begin{bmatrix} x_a - \mu_a \\ x_b - \mu_b \end{bmatrix}\right), \tag{21}$$

where Z' is a normalization constant that we used to absorb factors not depending on x_b.

$$p\left(x_b|x_a\right) = \frac{1}{Z'} \exp\left(-\frac{1}{2}\begin{bmatrix} x_a - \mu_a \\ x_b - \mu_b \end{bmatrix}^T \begin{bmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{bmatrix} \begin{bmatrix} x_a - \mu_a \\ x_b - \mu_b \end{bmatrix}\right) \tag{22}$$

$$= \frac{1}{Z'} \exp\left(-\left[\frac{1}{2}\left(x_a - \mu_a\right)^T \Lambda_{aa}\left(x_a - \mu_a\right) + \frac{1}{2}\left(x_a - \mu_a\right)^T \Lambda_{ab}\left(x_b - \mu_b\right)\right.\right. \tag{23}$$

$$\left.\left.+\frac{1}{2}\left(x_b - \mu_b\right)^T \Lambda_{ba}\left(x_a - \mu_a\right) + \frac{1}{2}\left(x_b - \mu_b\right)^T \Lambda_{bb}\left(x_b - \mu_b\right)\right]\right). \tag{24}$$

Recall the "completing the square" argument

$$\frac{1}{2}z^T Az + b^T z + c = \frac{1}{2}\left(z + A^{-1}b\right)^T A\left(z + A^{-1}b\right) + c - \frac{1}{2}b^T A^{-1}b. \tag{25}$$

Let

$$z = x_b - \mu_b, \tag{26}$$

$$A = \Lambda_{bb}, \tag{27}$$

$$b = \Lambda_{ba}\left(x_a - \mu_a\right), \tag{28}$$

$$c = \frac{1}{2}\left(x_a - \mu_a\right)^T \Lambda_{aa}\left(x_a - \mu_a\right). \tag{29}$$

Then, it follows that the expression for $p\left(x_b|x_a\right)$ can be rewritten as

$$p\left(x_b|x_a\right)=\frac{1}{Z'}\exp\left(-\left[\frac{1}{2}\left(x_b-\mu_b+\Lambda_{bb}^{-1}\Lambda_{ba}\left(x_a-\mu_a\right)\right)^T\Lambda_{bb}\left(x_b-\mu_b+\Lambda_{bb}^{-1}\Lambda_{ba}\left(x_a-\mu_a\right)\right)\right.\right. \tag{30}$$

$$\left.\left.+\frac{1}{2}\left(x_a-\mu_a\right)^T\Lambda_{aa}\left(x_a-\mu_a\right)-\frac{1}{2}\left(x_a-\mu_a\right)^T\Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba}\left(x_a-\mu_a\right)\right]\right). \tag{31}$$

Absorbing the portion of the exponent which does not depend on x_b into the normalization constant, we have

$$p\left(x_b|x_a\right)=\frac{1}{Z''}\exp\left(-\frac{1}{2}\left(x_b-\mu_b+\Lambda_{bb}^{-1}\Lambda_{ba}\left(x_a-\mu_a\right)\right)^T\Lambda_{bb}\left(x_b-\mu_b+\Lambda_{bb}^{-1}\Lambda_{ba}\left(x_a-\mu_a\right)\right)\right). \tag{32}$$

Looking at the last form, $p\left(x_b|x_a\right)$ has the form of a Gaussian density with mean $\mu_b-\Lambda_{bb}^{-1}\Lambda_{ba}\left(x_a-\mu_a\right)$ and covariance matrix Λ_{bb}^{-1}. Recall our matrix identity,

$$\begin{bmatrix}\Sigma_{aa} & \Sigma_{ab}\\ \Sigma_{ba} & \Sigma_{bb}\end{bmatrix}=\begin{bmatrix}\left(\Lambda_{aa}-\Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba}\right)^{-1} & -\left(\Lambda_{aa}-\Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba}\right)^{-1}\Lambda_{ab}\Lambda_{bb}^{-1}\\ -\Lambda_{bb}^{-1}\Lambda_{ba}\left(\Lambda_{aa}-\Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba}\right)^{-1} & \left(\Lambda_{bb}-\Lambda_{ba}\Lambda_{aa}^{-1}\Lambda_{ab}\right)^{-1}\end{bmatrix} \tag{33}$$

From this, it follows that

$$\mu_{b|a}=\mu_b-\Lambda_{bb}^{-1}\Lambda_{ba}\left(x_a-\mu_a\right)=\mu_b+\Sigma_{ba}\Sigma_{aa}^{-1}\left(x_a-\mu_a\right). \tag{34}$$

Conversely, we can also apply our matrix identity to obtain:

$$\begin{bmatrix}\Lambda_{aa} & \Lambda_{ab}\\ \Lambda_{ba} & \Lambda_{bb}\end{bmatrix}=\begin{bmatrix}\left(\Sigma_{aa}-\Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}\right)^{-1} & -\left(\Sigma_{aa}-\Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}\right)^{-1}\Sigma_{ab}\Sigma_{bb}^{-1}\\ -\Sigma_{bb}^{-1}\Sigma_{ba}\left(\Sigma_{aa}-\Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}\right)^{-1} & \left(\Sigma_{bb}-\Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab}\right)^{-1},\end{bmatrix} \tag{35}$$

from which it follows that

$$\Sigma_{b|a}=\Lambda_{bb}^{-1}=\Sigma_{bb}-\Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab}. \tag{36}$$

□