CS4487: Home Assignment №1

Department of Computer Science
City University of Hong Kong

Due on September 25, 2019, 7pm

Exercise 1

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

a) [1 points]. 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)?

b) [1 points]. 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 $\mathbb{E}[\mathbb{I}[X = 8]]$.

Exercise 1

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

a) [1 points]. 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)?

Use face number as random variable.

$P(\text{win}) = P(X=2) + P(X=4) + P(X=6)$

$= 0.1 + 0.2 + 0 = 0.3 < 0.5$

It's worse than a fair die

b) [1 points]. 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 $\mathbb{E}[\mathbb{I}[X = 8]]$.

$E[\mathbb{I}[X=8]]$

$= \sum_{x \in X} p(X=x)\, \mathbb{I}[X=8]$

$= p_3 \mathbb{I}[3=8] + p_8 \mathbb{I}[8=8] + p_9 \mathbb{I}[9=8]$

$= p_8.$

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:

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

$H(X,Y) = -\sum_{x \in X} \sum_{y \in Y} P(X=x, Y=y) \log_2 P(X=x, Y=y)$

$= -\sum_{x \in X} \sum_{y \in Y} P(X=x|Y=y) \cdot P(Y=y) \log_2 [P(X=x|Y=y) P(Y=y)]$

$= -\sum_{x \in X} \sum_{y \in Y} P(X=x|Y=y) P(Y=y) \log_2 P(X=x|Y=y)$
$- \sum_{x \in X} \sum_{y \in Y} P(X=x|Y=y) P(Y=y) \log_2 P(Y=y)$

$= -[\sum_{x \in X} P(X=x|Y=y) \cdot \log_2 P(X=x|Y=y)] \cdot [\sum_{y \in Y} P(Y=y)]$
$-[\sum_{y \in Y} P(Y=y) \log_2 P(Y=y)] \cdot [\sum_{x \in X} P(X=x|Y=y)]$

$= -[\sum_{x \in X} P(X=x|Y=y) \cdot \log_2 P(X=x|Y=y)]$
$-[\sum_{y \in Y} P(Y=y) \log_2 P(Y=y)].$

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

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

for all $x \in \mathcal{X}$ 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.

When $p(X=x, Y=y) = P(X=x) P(Y=y)$

$I(X,Y) = \sum_{x \in X} \sum_{y \in Y} P(X=x, Y=y) \log \left(\frac{P(X=x, Y=y)}{P(X=x) P(Y=y)} \right)$

$= \sum_{x \in X} \sum_{y \in Y} P(X=x, Y=y) \cdot 0$

$= 0.$

Exercise 2

[5 points]. Given a training set $\mathcal{D} = \{(x^{(i)}, y^{(i)}), i = 1, \ldots, m\}$, where $x^{(i)} \in \mathbb{R}^n$ and $y^{(i)} \in \mathcal{Y}$, derive the maximum likelihood estimates of the naive Bayes for real valued $x_j^{(i)}$ modeled with a univariate Gaussian distribution, *i.e.*,

$$p(X_j = x_j | Y = y) = \frac{1}{\sqrt{2\pi\sigma_{j|y}^2}} \exp\left(-\frac{(x_j - \mu_{j|y})^2}{2\sigma_{j|y}^2}\right).$$

$P(D; \phi, \mu, \sigma^2)$

$= \prod_{i=1}^{m} P(x^{(i)}, y^{(i)}; \phi, \mu, \sigma^2)$

$= \prod_{i=1}^{m} \left(P(y^{(i)}; \phi) \cdot P(x^{(i)} | y^{(i)}; \mu, \sigma^2) \right)$

$= \prod_{i=1}^{m} \left(P(y^{(i)}; \phi) \prod_{j=1}^{n} P(x_j^{(i)} | y^{(i)}; \mu_{j|y}, \sigma_{j|y}^2) \right)$

To maximise P, it is correct to maximise $\log P$

$\ell(\phi, \mu, \sigma^2) = \log L(\phi, \mu, \sigma^2)$

$= \log \prod_{i=1}^{m} \left(P(y^{(i)}; \phi) \prod_{j=1}^{n} P(x_j^{(i)} | y^{(i)}; \mu_{j|y}, \sigma_{j|y}^2) \right)$

$= \sum_{i=1}^{m} \log P(y^{(i)}; \phi) + \sum_{i=1}^{m} \sum_{j=1}^{n} \log P(x_j^{(i)} | y^{(i)}; \mu_{j|y}, \sigma_{j|y}^2)$

$= \underbrace{\sum_{i=1}^{m} \sum_{y \in Y} \mathbb{I}[y^{(i)} = y] \log \phi_y}_{①} + \underbrace{\sum_{i=1}^{m} \sum_{j=1}^{n} \log P(x_j^{(i)} | y^{(i)}; \mu_{j|y}, \sigma_{j|y}^2)}_{②}$

Optimise ①: Constraint: $\sum_{y \in Y} \phi_y = 1$

$J(\phi_y) = \sum_{i=1}^{m} \sum_{y \in Y} \mathbb{I}[y^{(i)} = y] \log \phi_y + \lambda \left(\sum_{y \in Y} \phi_y - 1 \right)$

$\frac{\partial J}{\partial \phi_y} = \frac{\sum_{i=1}^{m} \mathbb{I}[y^{(i)} = y]}{\phi_y} + \lambda = 0$

$\phi_y = \frac{\sum_{i=1}^{m} \mathbb{I}[y^{(i)} = y]}{-\lambda}$

$\sum_{y \in Y} \phi_y = \frac{\sum_{y \in Y} \sum_{i=1}^{m} \mathbb{I}[y^{(i)} = y]}{-\lambda} = 1$

$-\lambda = \sum_{i=1}^{m} \sum_{y \in Y} \mathbb{I}[y^{(i)} = y]$

$-\lambda = m$

$\phi_y = \frac{\sum_{i=1}^{m} \mathbb{I}[y^{(i)} = y]}{m}$

Optimize ①. $\sum_{i=1}^{m}\sum_{j=1}^{n} \log p(x_j^{(i)} | y^{(i)}; \mu_{jy}, \sigma_{jy}^2)$

Pick out all training samples with class label y by adding the indicator function $\mathbb{1}[y^{(i)}=y]$

$$J(\mu_{jy}, \sigma_{jy}^2) = \sum_{i=1}^{m} \mathbb{1}[y^{(i)}=y] \sum_{x \in X_j} \log\left(\frac{1}{\sqrt{2\pi\sigma_{jy}^2}} e^{-\frac{(x_j^{(i)}-\mu_{jy})^2}{2\sigma_{jy}^2}} \right)$$

$$= \sum_{i=1}^{m} \mathbb{1}[y^{(i)}=y] \sum_{x \in X_j} \left[\log\left(\frac{1}{\sqrt{2\pi\sigma_{jy}^2}} \right) - \frac{(x_j^{(i)}-\mu_{jy})^2}{2\sigma_{jy}^2} \right]$$

$$l = \sum_{i=1}^{m} \mathbb{1}[y^{(i)}=y] \sum_{x \in X_j} \frac{1}{2} \left[-\log 2\pi - \log\sigma^2 - \frac{(x_j^{(i)}-\mu_{jy})^2}{\sigma_{jy}^2} \right]$$

$$\frac{\partial l}{\partial \mu_{jy}} = -\sum_{i=1}^{m} \mathbb{1}[y^{(i)}=y] \frac{x_j^{(i)}-\mu_{jy}}{\sigma_{jy}^2} = 0$$

$$\sum_{i=1}^{m} \mathbb{1}[y^{(i)}=y] x_j^{(i)} = \sum_{i=1}^{m} \mathbb{1}[y^{(i)}=y] \mu_{jy} = \mu_{jy}\left[\sum_{i=1}^{m} \mathbb{1}[y^{(i)}=y]\right]$$

$$\mu_{jy} = \frac{\sum_{i=1}^{m} \mathbb{1}[y^{(i)}=y] x_j^{(i)}}{\sum_{i=1}^{m} \mathbb{1}[y^{(i)}=y]}$$

$$\frac{\partial l}{\partial \sigma_{jy}^2} = -\frac{1}{2}\sum_{i=1}^{m} \mathbb{1}[y^{(i)}=y]\left[\frac{1}{\sigma_{jy}^2} - \frac{(x_j^{(i)}-\mu_{jy})^2}{(\sigma_{jy}^2)^2}\right] = 0$$

$$\frac{1}{2}\sum_{i=1}^{n} \mathbb{1}[y^{(i)}=y]\left[\frac{(x_j^{(i)}-\mu_{jy})^2}{(\sigma_{jy}^2)^2} - \frac{1}{\sigma_{jy}^2}\right] = 0$$

$$\sum_{i=1}^{m} \mathbb{1}[y^{(i)}=y]\left[(x_j^{(i)}-\mu_{jy})^2 - \sigma_{jy}^2\right] = 0$$

$$\sum_{i=1}^{m} \mathbb{1}[y^{(i)}=y](x_j^{(i)}-\mu_{jy})^2 = \sum_{i=1}^{m} \mathbb{1}[y^{(i)}=y] \cdot \sigma_{jy}^2$$

$$\sigma_{jy}^2 = \frac{\sum_{i=1}^{m} \mathbb{1}[y^{(i)}=y](x_j^{(i)}-\mu_{jy})^2}{\sum_{i=1}^{m} \mathbb{1}[y^{(i)}=y]}$$

Exercise 3

[5 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

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

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

$P(Y=y|X=x;\phi,\mu,\Sigma)$ where Σ is diagonal

$$= \frac{P(X=x|Y=y;\phi,\mu,\Sigma) \cdot P(Y=y;\phi)}{P(X=x;\phi,\mu)}$$

$$P(Y=1|X=x;\phi,\mu,\Sigma)$$

$$= \frac{\phi_1 P(X=x|Y=1;\phi_1,\mu_{x|1},\Sigma_{x|1})}{\phi_0 P(X=x|Y=1;\phi_0,\mu_{x|0},\Sigma_{x|0}) + \phi_1 P(X=x|Y=1;\phi_1,\mu_{x|1},\Sigma_{x|1})}$$

$$\frac{P(Y=1|X=x)}{P(Y=0|X=x)} = \frac{\phi_1 P(X=x|Y=1;\phi_1,\mu_{x|1},\Sigma_{x|1})}{\phi_0 P(X=x|Y=1;\phi_0,\mu_{x|0},\Sigma_{x|0})}$$

since Σ is showed.

$$= \frac{\phi_1}{\phi_0} \cdot \frac{\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_{i|y=0}}}}{\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_{i|y=1}}}} \cdot \exp\left(-(x_i-\mu_1)^T\Sigma^{-1}(x_i-\mu_1)/2 + (x_i-\mu_0)^T\Sigma^{-1}(x_i-\mu_0)/2\right)$$

since $\Sigma_{y=0} = \Sigma_{y=1}$.

$$\frac{P(Y=1|X=x)}{P(Y=0|X=x)} = \frac{\phi_1}{\phi_0} \cdot \exp\left(-(x_i-\mu_1)^T\Sigma^{-1}(x_i-\mu_1)/2 + (x_i-\mu_0)^T\Sigma^{-1}(x_i-\mu_0)/2\right)$$

$$P(Y=0|X=x) = \frac{P(Y=1|X=x) \cdot \phi_0}{\phi_1 \exp\left(-(x_i-\mu_1)^T\Sigma^{-1}(x_i-\mu_1)/2 + (x_i-\mu_0)^T\Sigma^{-1}(x_i-\mu_0)/2\right)}$$

$$P(Y=1|X=x) + P(Y=0|X=x)$$

$$= P(Y=1|X=x) \cdot \left(1 + \frac{1}{\frac{\phi_1}{\phi_0}\exp\left(-(x_i-\mu_1)^T\Sigma^{-1}(x_i-\mu_1)/2 + (x_i-\mu_0)^T\Sigma^{-1}(x_i-\mu_0)/2\right)} \right)$$

$$= P(Y=1|X=x) \cdot \left(1 + \frac{\frac{\phi_1}{\phi_0}\exp\left(-(x_i-\mu_1)^T\Sigma^{-1}(x_i-\mu_1)/2 + (x_i-\mu_0)^T\Sigma^{-1}(x_i-\mu_0)/2\right)}{\frac{\phi_1}{\phi_0}\exp\left(-(x_i-\mu_1)^T\Sigma^{-1}(x_i-\mu_1)/2 + (x_i-\mu_0)^T\Sigma^{-1}(x_i-\mu_0)/2\right)} \right)$$

$$= P(Y=1|X=x) \cdot \left(\frac{\frac{\phi_0}{\phi_1}\exp\left((x_i-\mu_1)^T\Sigma^{-1}(x_i-\mu_1)/2 - (x_i-\mu_0)^T\Sigma^{-1}(x_i-\mu_0)/2\right)}{1} + 1 \right)$$

$$= 1$$

$$P(Y=1|X=x) = \frac{1}{1 + \frac{\phi_0}{\phi_1}\exp\left((x-\mu_1)^T\Sigma^{-1}(x-\mu_1) - (x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\right)}$$

$$= \frac{1}{1 + \exp\left(\ln\left(\frac{\phi_0}{\phi_1}\right) + \frac{1}{2}\left(x^T\Sigma^{-1}(x-\mu_1) - \mu_1^T\Sigma^{-1}(x-\mu_1) - x^T\Sigma^{-1}(x-\mu_0) + \mu_0^T\Sigma^{-1}(x-\mu_0)\right)\right)}$$

$$= \frac{1}{1 + \exp\left(\ln\frac{\phi_0}{\phi_1} + \frac{1}{2}\mu_{y=1}^T\Sigma^{-1}\mu_{y=1} - \frac{1}{2}\mu_{y=0}^T\Sigma^{-1}\mu_{y=0} + (\mu_0-\mu_1)^T\Sigma^{-1}x\right.}$$

$$= \frac{1}{1 + \exp\left[-\left(-(\mu_0-\mu_1)^T\Sigma^{-1}x - \left(\ln\frac{\phi_0}{\phi_1} + \frac{1}{2}(\mu_{y=1}^T\Sigma^{-1}\mu_{y=1} - \mu\right.\right.\right.}$$

where we set $x_0 = 1$

therefore: $\frac{1}{2}(\mu_{y=1}^T\Sigma^{-1}\mu_{y=1} - \mu_{y=0}^T\Sigma^{-1}\mu_{y=0}) + \ln\frac{\phi_0}{\phi_1}$ is produced by bias

posterior of the linear discriminant analysis admits the sigmoid form: $\frac{1}{1+e^{-\theta x}}$

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:

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

d) [1 points]. 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.

Exercise 2

[5 points]. Given a training set $\mathcal{D} = \{(x^{(i)}, y^{(i)}), i = 1,\ldots,m\}$, where $x^{(i)} \in \mathbb{R}^n$ and $y^{(i)} \in \mathcal{Y}$, derive the maximum likelihood estimates of the naive Bayes for real valued $x_j^{(i)}$ modeled with a univariate Gaussian distribution, *i.e.*,

$$p(X_j = x_j|Y=y) = \frac{1}{\sqrt{2\pi\sigma_{j|y}^2}}\exp\left(-\frac{(x_j-\mu_{j|y})^2}{2\sigma_{j|y}^2}\right).$$

Exercise 3

[5 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

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

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