# Homework 1

Maxwell Aladago

February 12, 2018

## 1 q1

Given two random variables x and y

$$cov(x, y) = \mathbb{E}[(x - \mathbb{E}[x])(y - \mathbb{E}[y])]$$

Expanding and distributing the expectation

$$cov(x, y) = \mathbb{E}_{x,y}[xy] - \mathbb{E}[x\mathbb{E}[y]] - \mathbb{E}[y\mathbb{E}[x]] + \mathbb{E}[\mathbb{E}[x]\mathbb{E}[y]]$$

Since the E[E[D]] = E[D] for all Random variables D

$$\begin{aligned} cov(x, y) &= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y] - \mathbb{E}[y]\mathbb{E}[x] + \mathbb{E}[x]\mathbb{E}[y] \\ &= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y] \end{aligned} \tag{1.1}$$

From the definition, if x and y are independent:

**Lemma 1.1**

$$\mathbb{E}[xy] = E[x][y]$$

Hence,

$$cov(x, y) = E[x][y] - E[x][y] = 0$$

## 2 q2

Let *B* be the random variable denoting the box chosen and *F* be the random variable denoting the fruit picked. Also let

$$r, g, b$$

represent the red box, greed box and blue box respectively. Thus the prior probabilities of selecting the red box, the green box and the blue box are defined respectively:

$$p(B = r) = 0.2 = \frac{1}{5}$$

$$p(B = g) = 0.5 = \frac{1}{2}$$

$$p(B = b) = 0.3 = \frac{3}{10}$$

## 2.1 q2 a. Probability of selecting an apple

The marginal probability of selecting an apple is given by

$$p(F = a)$$

Applying the product and sum rules:

$$p(F = a) = p(F = a|B = r).p(B = r) + p(F = a|B = g).p(B = g) + p(F = a|B = b).p(B = b) \quad (2.1)$$

Also from the question and using the definition of probability, the conditional probabilities in Eq.2.1 above can be specified as follows

$$p(F = a|B = r) = \frac{3}{10}$$
$$p(F = a|B = g) = \frac{3}{10}$$
$$p(F = a|B = b) = \frac{1}{3}$$

By substituting the respective probabilities, Eq. 2.1 evaluates to:

$$p(F = a) = (\frac{3}{10} \times \frac{1}{5}) + (\frac{3}{10} \times \frac{1}{2}) + (\frac{1}{3} \times \frac{3}{10})$$

**Solution 2.1**

$$p(F = a) = \frac{3}{50} + \frac{3}{20} + \frac{1}{10} = \frac{31}{100} = \underline{\underline{\boldsymbol{0.31}}}$$

## 2.2 q2 b. Probability that a fruit was selected from the green box given it is an orange

The posterior probability of selecting the green box given that an orange is chosen is expressed as

$$p(B = g|F = o)$$

According to *Bayes' theorem*

**Definition 2.1**

$$p(B = g|F = o) = \frac{p(F = o|B = g).p(B = g)}{p(F = o)}$$

Where:

$$p(F = o|B = g) = \frac{3}{10} \quad \text{(i)}$$
$$p(B = g) = \frac{1}{2} \quad \text{(ii)}$$

$$p(F = o) = p(F = o|B = r).p(B = r) + p(F = o|B = g).p(B = g) + p(F = o|B = b).p(B = b) \quad (2.2)$$

The conditional probabilities in Eq. 2.2 can be specified as:

$$p(F = o|B = r) = \frac{4}{10}$$

$$p(F = o|B = g) = \frac{3}{10}$$

$$p(F = o|B = b) = \frac{2}{3}$$

Hence,

$$p(F = o) = (\frac{4}{10} \times \frac{1}{5}) + (\frac{3}{10} \times \frac{1}{2}) + (\frac{2}{3} \times \frac{3}{10}) = \frac{43}{100} \tag{iii}$$

Substituting (i), (ii) and (iii) into Def. 2.1,

**Solution 2.2**

$$p(B = g|F = o) = \frac{\frac{3}{10} \times \frac{1}{2}}{\frac{43}{100}} = \frac{15}{43} \approx \underline{\underline{\boldsymbol{0.35}}}$$

# 3 q3

## 3.1 q3a

Definition of terms:

$$p(head) = p(c = 1; \mu) = \mu$$

If

$$c = \{1, 0\}$$

is a random variable of the results of the flip, then the probability distribution over c can be expressed as

$$P(c; \mu) = \mu^c (1 - \mu)^{1-c}$$

If H is the number of times c = 1 in a sample data D = $\{c^{(1)}, c^{(2)}, c^{(3)}, ..., c^{(m)}\}$, and since the flips of the coin are independent, the likelihood function is given as:

$$\begin{aligned} p(D; \mu) &= \prod_{i=1}^{m} \mu^{c^i} (1 - \mu)^{1-c^i} \\ &= \prod_{i=1}^{H} \mu^{c^i} \prod_{i=1}^{m-H} (1 - \mu)^{1-c^i} \end{aligned}$$

**Solution 3.1**

$$L(\mu) = p(D; \mu) = \mu^H (1 - \mu)^{m-H}$$

## 3.2 q3b: Deriving the parameter which maximizes the likelihood

Writing the likelihood function in Soln. 3.1 above,

$$L(\mu) = \mu^H(1-\mu)^{m-H}$$

Let

$$l(\mu) = logL(\mu)$$

$$\begin{aligned} l(\mu) &= \log(\mu^H(1-\mu)^{m-H} \\ &= \log\mu^H + \log(1-\mu)^{m-H} \\ &= H\log\mu + m\log(1-\mu) - H\log(1-\mu) \end{aligned} \tag{3.1}$$

Taking $\frac{\delta l}{\delta \mu}$ of Eq. 3.1,

$$\begin{aligned} \frac{\delta l}{\delta \mu} &= \frac{\delta}{\delta \mu}[H\log\mu + m\log(1-\mu) - H\log(1-\mu)] \\ &= \frac{H}{\mu} - \frac{m}{1-\mu} + \frac{H}{1-\mu} \\ &= \frac{H(1-\mu) - m\mu + H\mu}{\mu(1-\mu)} \end{aligned}$$

At the optimal point, $\frac{\delta l}{\delta \mu} = 0$.
Hence,

$$0 = \frac{H(1-\mu) - m\mu + H\mu}{\mu(1-\mu)}$$
$$H - H\mu - m\mu + H\mu = 0$$

**Solution 3.2**

$$\mu_{ML} = \frac{H}{m}$$

Thus, the parameter maximizing the likelihood is the sample proportion of heads in the data.

## 3.3 q3e

The prior distribution of $\mu$ is given by

$$p(\mu; a) = \frac{1}{Z}\mu^{a-1}(1-\mu)^{a-1}$$

Where a parameter governing the distribution and Z is a constant

The posterior distribution of $\mu$ is

$$\begin{aligned} p(\mu|D,a) &= \frac{p(D|\mu) \times p(\mu; a)}{p(D)} \\ &= \frac{(\mu^H(1-\mu)^{m-H})(\mu^{a-1}(1-\mu)^{a-1})}{Z.p(D)} \end{aligned}$$

Since p(D), the evidence, is a constant,

$$\begin{aligned} p(\mu|D,a) &\propto \frac{1}{Z}(\mu^H(1-\mu)^{m-H}).(\mu^{a-1}(1-\mu)^{a-1}) \\ &\propto \frac{1}{Z}\{\mu^{H+a-1}(1-\mu)^{m+a-H-1}\} \\ &= \frac{1}{Z}\{\mu^{H+a-1}(1-\mu)^{m-H+a-1}\} \end{aligned} \tag{3.2}$$

Estimating the MAP of Eq. 3.2 and dropping the constant Z
Let

$$\begin{aligned} l(\mu) &= log(\mu^{H+a-1}(1-\mu)^{m-H+a-1}) \\ &= log\mu^{H+a-1} + log((1-\mu)^{m-H+a-1}) \\ &= (H+a-1)log\mu + (m-H+a-1)log(1-\mu) \end{aligned}$$

Taking $\frac{\delta l}{\delta \mu}$ of $l(\mu)$

$$\begin{aligned} \frac{\delta l}{\delta \mu} &= \frac{\delta}{\delta \mu}[(H+a-1)log\mu + (m-H+a-1)log(1-\mu)] \\ &= \frac{H+a-1}{\mu} - \frac{m-H+a-1}{1-\mu} \\ &= \frac{(H+a-1)(1-\mu) - \mu(m-H+a-1)}{\mu(1-\mu)} \\ &= \frac{H+a-1-2a\mu+2\mu-m\mu}{\mu(1-\mu)} \end{aligned}$$

For MAP estimate of $\mu$, $\frac{\delta l}{\delta \mu} = 0$
Therefore,

$$\frac{H+a-1-2a\mu+2\mu-m\mu}{\mu(1-\mu)} = 0$$

$$H+a-1 = \mu(m+2a-2)$$

**Solution 3.3**

$$\mu_{MAP} = \frac{H+a-1}{m+2(a-1)}$$

## 3.4 q3g

From the MAP estimate in Soln.3.3, the parameter a can be interpreted as the proportion of training examples incorporated in the prior distribution.

# 4 q4

## 4.1 q4c

The model $b^l(x)$ had underfitting for values of $\lambda = 10^3$, $\lambda = 10^5$ and $\lambda = 10^7$. On the other hand, $b^q(x)$ had underfitting for $\lambda = 10^7$.

## 4.2 q4d

For the $b^l(x)$, did not produce any overfitting for the given values of $\lambda$. However, the model $b^q(x)$ produced significant overfitting for $\lambda = 10^{-5}$, $\lambda = 10^{-3}$ and $\lambda = 10^{-1}$. It also overfit for $\lambda = 10^1$ and $\lambda = 10^3$. The magnitude of overfitting decreased from $\lambda = 10^{-5}$ through $\lambda = 10^{^3}$

## 4.3 q4e

The feature vector $b^q(x)$ tend to produce more overfitting compared to $b^l(x)$. Actually, $b^q(x)$ is too simple to produce any overfitting for the given values of $\lambda$ The reason is that $b^q(x)$ being a quadratic in the features produces a more complex model than $b^l(x)$. For very small values of $\lambda$ such as $\lambda = 10^{(-5)}$, the contribution of the smoothness term $\frac{\lambda}{2}||\theta||^2$ is minimal which then allows the model to over-fit the training data.

## 4.4 q4g

Yes, the cross validation is a good performance of test set. The reason is because every sample is used for both training and validation in cross validation. This enables the model to generalize well hence a good indication of performance on test set. This explanation has been concretized in [P, 4f][1] where the training and test errors are highly correlated for the linear model for instance.

[1]The programming test file q4f.py. Running this file produces plots of training and testing errors.