Ruth Nyakio Karimi

Machine Learning

H. Assignment 1

Tuesday 13th Feb 2018

(1) Given random variable X & Y. Show covariance is zero ie $cov(X,y) = 0$.

If X & Y are independent, then they are incorreelated. $cov(X,y) =$

Given:

$$cov(X,y) = E((X, EX)(Y - EY))$$
$$= E(X - (Y - EY) - EX(Y - EY))$$
$$= E(XY - XEY - YEX + EXEY$$
$$= E(XY) - EXEY - EYEX + EXEY$$

Therefore

$$cov(X,y) = E(XY) - EXEY$$

For independent variable X & Y

$$E(XY) = E(X) \cdot E(Y)$$
$$= EX \cdot EY$$

Hence

$$= EX \cdot EY \cdot EXEY$$
$$= 0$$

Therefore this proves that given one-dimensional random Variables X & Y and they are independent, their covariance is zero

$$cov(X,Y) = 0$$

6.

Question 2

	Apples	Oranges	limes	Total	P
red (r)	3	4	3	10	0.2
green (g)	3	3	4	10	0.5
blue (b)	1	2	0	3	0.3

(1) Probability of selecting an apple

$p(a) = p(r \text{ and } a) \text{ or } p(g \text{ and } a) \text{ or } p(b \text{ and } a)$

$= p(r \text{ and } a) = p(r) \text{ and } p(a)$

$= 0.2 \times 0.3$

$= 0.06$

$= p(g \text{ and } a) = p(g) \cdot p(a)$

$= 0.5 \times 0.3$

$= 0.15$

$= p(b \text{ and } a) = p(b) \text{ and } p(a)$

$= 0.3 \times 1/3$

$= 3/10 \times 1/3 = 1/10$

$= 0.1$

Therefore $p(a) = 0.06 + 0.15 + 0.1$

$= 0.31$

Therefore, the probability of selecting an apple is $\underline{\underline{0.31}}$

6.

(ii) Probability than an orange came from the green box.

$$P(g|o) = \frac{p(o|g) \cdot p(g)}{P(o)}$$ Bayes's rule

$= p(o|g) = 3/10 = 0.3$

$= p(g) = 0.5$

$= p(o) = p(o \text{ and } r)$ or $p(o \text{ and } b)$ or $p(o \text{ and } g)$

$p(o \text{ and } r) = 4/10 \times 2/10 = 0.08$

$p(o \text{ and } b) = 2/3 \times 3/10 = 0.2$

$p(o \text{ and } g) = 3/10 \times 5/10 = 0.15$

$\therefore p(o) = 0.08 + 0.2 + 0.15$

$= 0.43$

Hence $= \dfrac{0.3 \times 0.5}{0.43}$

$= \dfrac{0.15}{0.43}$

$p(g|o) = 0.35$

Therefore; probability that a selected fruit orange came from the green box is $\boxed{0.35}$

Question 3

(a) Writing the likelihood function

Given $D = \{c^{(i)} \ldots\ldots c^{(m)}\}$

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

m times flip and

$c^{(i)}$ denoting i^{th} flip then:

Estimate μ

$$p(D; \mu) = \prod_{i=1}^{m} p(c^{(i)}; \mu) \quad \text{— notation.}$$

$$= \prod_{i=1}^{m} \left(\mu^{c(i)} (1-\mu)^{1-c(i)}\right)$$

$$= \mu^{c(1)} (1-\mu)^{1-c(1)} \times \mu^{c(2)} (1-\mu)^{1-c(2)} \cdots$$

$$\cdots \mu^{c(m)} (1-\mu)^{1-c(m)}$$

$$= \mu^{\sum_{i=1}^{m} c^{(i)}} (1-\mu)^{m - \sum_{i=1}^{m} c(i)}$$

given $\sum_{i=1}^{m} c^{(i)}$ can be rewritten as H

hence

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

Therefore the likelihood function or the probability of data D is:

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

(b) Deriving Parameter μ using Max. Likelihood

- Calculate derivative of $L(D, \mu)$ with respect μ
- Set derivative equal to zero (0)
- Solve the resulting equation for μ.

Given $p(D^{(i)}; \mu)$

$$= \log \left[\prod_{i=1}^{m} p(c_j^{(i)}; \mu) \right]$$

$$= \sum_{i=1}^{m} \log p(c_j^{(i)}; \mu)$$

Given distribution $p(c; \mu) = \mu^{c} (1-\mu)^{1-c}$

$$= \sum_{i=1}^{m} \left[\log (\mu^{c(i)}) (1-\mu)^{1-c(i)} \right]$$

$$= \sum_{i=1}^{m} \left[\log (\mu)^{c(i)} + \log (\mu-1)^{1-c(i)} \right]$$

given $\log a^b = b \log a$

$$= \sum_{i=1}^{m} \left[c^{(i)} \log \mu + 1 - c^{(i)} \log (1-\mu) \right]$$

$$= \sum_{i=1}^{m} c^{(i)} \log \mu + \sum_{i=1}^{m} (1-c^{(i)}) \log (1-\mu)$$

$$= \sum_{i=1}^{m} c^{(i)} \log \mu + \left(m - \sum_{i=1}^{m} c^{(i)} \right) \log (1-\mu)$$

- derivatives of log

$$\frac{d}{d\mu} \operatorname{Log} \mu = 1/\mu \quad \text{and}$$

$$\frac{d}{d\mu} \log (1-\mu) = \frac{1}{\mu - 1}$$

Hence

$$= \frac{\sum_{i=1}^{m} c^{(i)}}{\mu} + \frac{\left(m - \sum_{i=1}^{m} c^{(i)}\right)}{\mu - 1} \equiv 0$$

- Multiply both sides by $\mu(\mu-1)$

$$= (\mu-1)\left(\sum_{i=1}^{m} c^{(i)}\right) + \left(m - \sum_{i=1} c^{(i)}\right)\mu \equiv 0$$

$$= \cancel{\mu\left(\sum_{i=1}^{m} c^{(i)}\right)} - \sum_{i=1}^{m} c^{(i)} + \mu m - \cancel{\mu \sum_{i=1}^{m} c^{(i)}} \equiv 0$$

$$= \mu m - \sum_{i=1}^{m} c^{(i)} \equiv 0$$

$$= \frac{\mu m}{m} = \frac{\sum_{i=1}^{m} c^{(i)}}{m}$$

$$= \mu = 1/m \sum_{i=1}^{m} c^{(i)}$$

given $\sum_{i=1}^{m} c^{(i)}$ can be written as H

$$\mu = H/m$$

Therefore $\boxed{\mu_{ML} = H/m}$ or $\boxed{\sum_{i=1}^{m} c^{(i)}}$

(3e) Deriving the analytical expression of parameter μ using MAP estimation.

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}$$

$$p(\theta|D) = \frac{p(D|\theta)\, p(\theta)}{p(D)}$$

$$\theta^* = \arg\max\, p(\theta/D) \propto \underbrace{p(D|\theta)\, p(\theta)}_{\varepsilon}$$

Given:

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

Prior $\quad p(\mu;a) = \frac{1}{2} \mu^{a-1} (1-\mu)^{a-1}$

then:

$$\mu^* = \mu^H (1-\mu)^{m-H} \cdot \frac{1}{2} \mu^{a-1} (1-\mu)^{a-1}$$

where $\mu^* = \arg\max_{\mu}\, p(\mu|D) \propto p(D|\mu)\, p(\mu)$

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

$$= \log\left[\frac{1}{2} \mu^{H+a-1} \cdot (1-\mu)^{m-H+(a-1)}\right]$$

$$= \frac{1}{2}\left[\log(\mu^{H+a-1} + \log(1-\mu)^{m-H+(a-1)}\right]$$

Take derivatives of log

$$= \frac{1}{2}\left[\frac{H+(a-1)}{\mu} + \frac{m-H+(a-1)}{\mu-1}\right] \equiv 0$$

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

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

$$=\mu-1\,(H+a-1)+\mu\,[m-H+a-1]=0$$

$$=H\mu-H+\mu a-\mu-a+1+(\mu m-\mu H+\mu a-\mu)=0$$

$$=-H-a+1+H\mu a+\mu a-\mu+\mu m-\mu H+\mu a-\mu=0$$

$$\cancel{=H-a+1=-H\mu a+\mu a-\mu-(\mu m-\mu H+\mu a-\mu)}$$

$$=\quad H+a-1=\mu\,[H+a-1+(m-H+a-1)]$$

$$\hat{\mu}_{MAP}=\frac{H+a-1}{(m-H)+(a-1)+H+a-1}$$

$$=\frac{H+a-1}{m+2a-2}$$

Therefore the analytical expression of parameter μ using MAP is

$$\hat{\mu}_{MAP}=\frac{H+a-1}{m+2a-2}$$

3f

In terms of training examples, parameter a can be interpreted as the:

Imaginary examples where the larger the a is, the more confident we are about the prior and the less the a is, the less confident or sure we are about the prior.

Question 4

(4c) Are there any values of λ producing underfitting?

Yes, there are values of λ producing underfitting.

$\lambda = 10^{3}, 10^{5}, 10^{7}$

Are there values producing overfitting. Yes there

(4d) are values producing overfitting.

$\lambda = 10^{-1}, 10^{-3}, 10^{-5}$

(4e) $b^2(x)$ produces more overfitting compared to $b^1(x)$ which is achieved by plotting the quadratic model $b^2(x)$ produces more overfitting compared to $b^1(x)$ as it tends to produce more reliance on the training data than $b^1(x)$.

(4g) Yes, cross-validation score a good predictor of performance on the test data. This is because it holds out part of the data for testing (x_test and Y_test) and tests data multiple times by providing different combinations of training and test data each time.