assignment 1

Jeppe Johansen

November 2017

1 Introduction

2 Probability Theory: Properties of expectations

Question 1

Proof that the relationship in equation 1 holds true.

$$E[X+Y] = E[X] + E[Y] \tag{1}$$

I know that expectation is defined as:

$$E[X] = \sum_{x \in X} x \cdot P(X = x) \tag{2}$$

From here I can deduce that:

$$E[X+Y] = \sum_{y \in Y} \sum_{x \in X} (x+y) P(X = x) P(Y = y) \tag{3}$$

where s in event in the sample space omega. Transforming expression in equation 3 I get:

$$E[X+Y] = \sum_{x \in X} x \cdot P(X = x) + \sum_{y \in Y} y \cdot P(Y = y) \tag{4}$$

And finally I have proofed the expression:

$$E[X+Y] = E[X] + E[Y] \tag{5}$$

Question 2

Proof that the relationship in equation 6 Given X & Y are independent.

$$E[XY] = E[X]E[Y] \tag{6}$$

Table 1: Joint Distribution

$X \mid Y$	0	1
0	$\frac{1}{4}$	0
1	$\frac{1}{2}$	0
2	0	$\frac{1}{4}$

I know that the definition of independence of random 2 random variables is: X & Y are independent if and only if $P(X = x) \cap P(Y = y) = P(X = x) \cdot P(Y = y)$. The expectation of the join distribution is:

$$E[XY] = \sum_{y \in Y} \sum_{x \in X} xyP(X = x, Y = y) \tag{7}$$

At this stage I assume independence and I get:

$$E[XY] = \sum_{y \in Y} \sum_{x \in X} xP(X = x) \cdot yP(Y = y) \tag{8}$$

$$\Leftrightarrow E[XY] = \sum_{x \in \Omega} xP(x) \cdot \sum_{y \in \Omega} yP(y) = E[x]E[y] \tag{9}$$

And the statement is proofed.

Question 3

Using the example and defining two random variables Z_1, Z_2 that are bernoulli distributed with parameter $p = 0.5$. With Z_1, Z_2 I define X, Y the following way:

$$X = Z_1 + Z_2, \qquad Y = Z_1 \cdot Z_2 \tag{10}$$

From here I calculate the join distribution and present it in the table 1:
First I calculate the expectation of the joint distribution:

$$E[XY] = (0 \cdot 0)\frac{1}{4} + (1 \cdot 1)\frac{1}{2} + (2 \cdot 1)\frac{1}{4} = \frac{1}{2} \tag{11}$$

$$E[X]E[Y] = (\underbrace{1\frac{1}{2} + 2\frac{1}{4}}_{E[X]}) \quad (\underbrace{1\frac{1}{4}}_{E[Y]}) = \frac{1}{4} \tag{12}$$

And I see that the product of the expectations of the random variables does not equal to the expectation of the product of the random variables i.e.

$$E[XY] \neq E[x]E[Y] \tag{13}$$

Question 4

Prooving that $E[E[X]] = E[X]$ goes the following way:

$$E[E[X]] = \sum_{x \in X} \left[\sum_{x \in X} xp(X = x) \right] p(X = x) \tag{14}$$

$$E[E[X]] = \sum_{x \in X} p(X = x) \left[\sum_{x \in X} xp(X = x) \right] \tag{15}$$

$$E[E[X]] = 1 \left[\sum_{x \in X} xp(X = x) \right] = \sum_{x \in X} xp(X = x) = E[X] \tag{16}$$

Question 5

$$Var[X] = E[(X - E[X])^2] = E[X^2 + E[X]^2 - 2E[X]] \tag{17}$$

which equals to:

$$= E[X^2] + E[E[X]^2] - E[X2E[X]] = E[X^2] + E[E[X]^2] - E[X]2E[X] \tag{18}$$

$$= E[X^2] + E[E[X]^2] - 2E[X]^2 = E[X^2] - E[X]^2 \tag{19}$$

3 Probability Theory: Complements of Events

Question 1

I know that from the axioms of probability spaces that $P(\Omega) = 1$, for all events drawn from the probability space $0 \leq P(event) \leq 1$ and that a sequence of countable disjoints events $event_1, event_2$ I can write: $P(\cup_{i>0}(event_i)) = \sum_{i>0} P(event_i)$.

Using that the set combined with the compliment of the set I have the full probability space.

$$\Omega = A \cup \bar{A} \tag{20}$$

From here I can conclude since the events A and $\bar{A}$:

$$P(\Omega) = P(A) + P(\bar{A}) = 1 \tag{21}$$

Rearranging the expression:

$$P(\Omega) - P(\bar{A}) = 1 - P(\bar{A}) = P(A) \tag{22}$$

Question 2

3.0.1 The probability of observing at least 1 tail:

I see the only way not to observe at least 1 tail is to to observe 0 tails. The probability of that event is:

$$P(\text{zero tails}) = \prod_{i=1}^{10} \left(\frac{1}{2}\right) = 0.000977 \tag{23}$$

This results implies:

$$P(\text{At least 1 tail}) = 1 - P(\text{zero tails}) = 1 - 0.000977 = 0.999023 \tag{24}$$

What is the probability to observe at least two tails?

I can deduce that if I have at least 2 tails, I can maximum get 1 tails. I have already computed the probability of getting zero tails. Next computing the probability of getting 1 and only 1 tails. This can happen in 10 different configurations. That the first flip of the coin is tails and all the others aren't OR the second flip of the coin is tails and all others are not, and so on and so forth.

$$P(\text{at least 1 tails}) = 10 * \prod_{i=1}^{10} \left(\frac{1}{2}\right) = 10 \cdot 0.000977 = 0.00977 \tag{25}$$

Which implies that the probability of at least two tails is:

$$\begin{aligned} P(\text{at least two tails}) &= 1 - P(1 \text{ tails}) + P(\text{zero tails}) && (26)\\ &= 1 - 0.00977 - 0.000977 && (27)\\ &= 0.989253 && (28) \end{aligned}$$

4 Digit Classification

The KNN-estimator is made as for each observations that I want to predict I do the following calculations: take the outer product of the features for given observation i and a vector of ones with that has dimensions ($1 \times \#$ observations in training set). calling the outer product M, I can compute this matrix from the training data so I get: $A = X_{training} - M$ since M and $X_{training}$ have the same dimensions. I can from here compute the euclidian distances. When all the euclidian distances is computed I can sort the labels after shortest distance. From here I let the algorithm choose the label for the observations I want to predict by doing a vote by majority.

Before doing any of the prediction I split the training data set into to data sets a validation set and a training data set. This means we end up with the

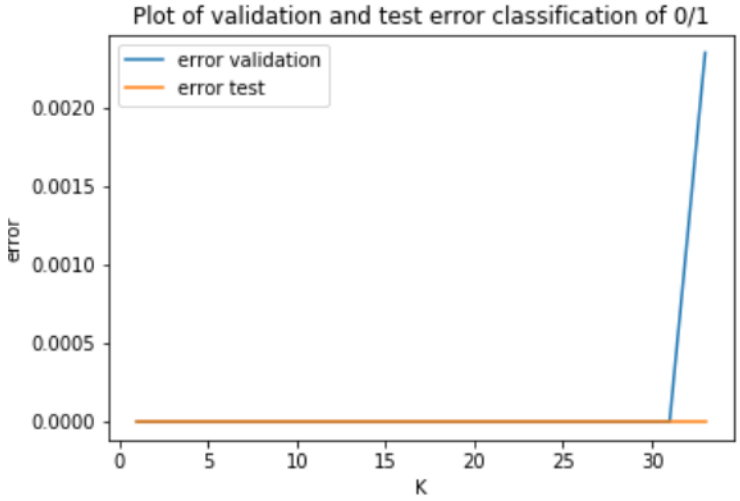

Figure 1:

following three datasets: training set of 8000 observations, validation set of 2000 observations, test set of 2000 observations observations.

Figure 1, 2, **??** shows the results of the classification where mean squared error is on the y-axis and the model complexity in the form of K is on the x-axis. All three classification is carried out with a high degree of succes but the algorithm seems to perform best on the $0/1$ classification, newt best on $0/8$ classifcation and poorest on $5/6$. For the case where classification is between $0/1$ it's unclear at which model complexity the model performs best since for $K \in \{1, 2, .., 31\}$ the error is 0 in both validation and test data set. For classification of $0/8$ the best K is probably in the range 5 to 10. For classification of $5/6$ the best value of K seems to be 3.

It's worth noting, that in this assignment good machine learning practice has been violated since the test data set has been used for inference of the parameter K (even if only indirectly).

5 Linear Regression

First we define the x and y variables with following code.

```
x = np.array(danwood.drop('y', axis=1))
y = np.array(danwood['y'])
```

No we compute the β parameters. The OLS estimate is calculated by the formula at 29:

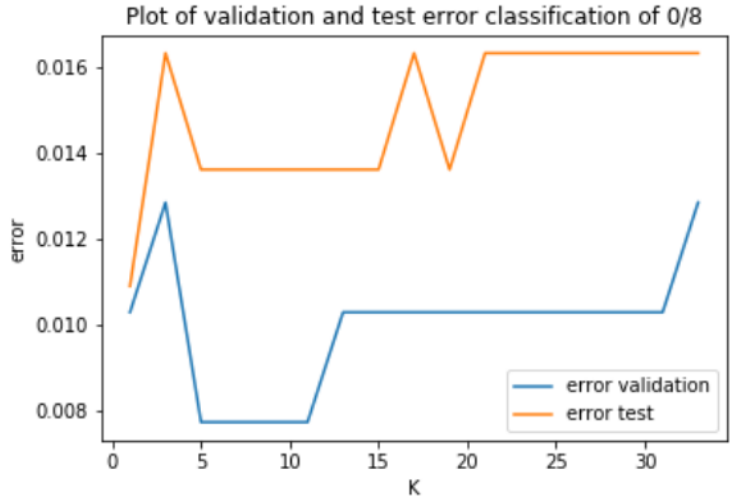

Figure 2:

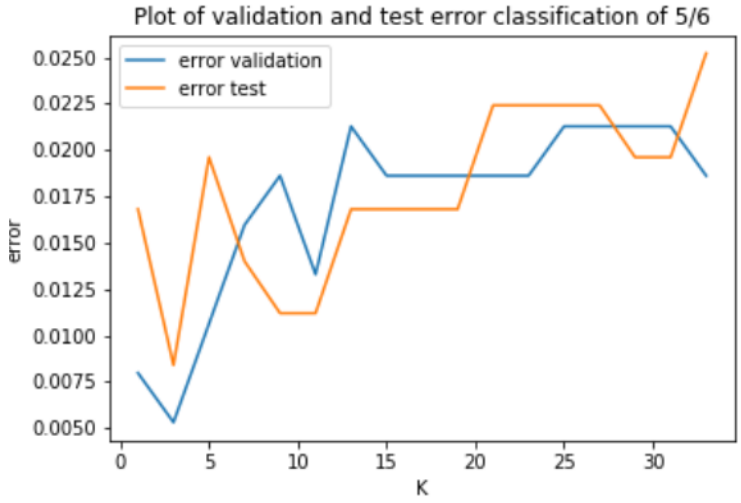

Figure 3:

$$\beta = (X^T X)^{-1}(X^T y) \tag{29}$$

The code below show how the calculations is done in the python, where numpy has been used for matrix operations.

```
xx = np.dot(x.T,x)
xy = np.dot(x.T,y)
xx_inv = np.linalg.inv(xx)

ols = np.dot(xx_inv,xy)
```

The result of the OLS is a intercept: -10.42, and β_1: 9.49. Furthermore we can calculate the MSE of the regression by equation 30, which is a measure of how well over model fit the data.

$$\text{Mean squared error} = \frac{1}{N}\sum_{i=1}^{N}(y_i - (9.49x_i - 10.42))^2 = 0.012 \tag{30}$$

$$var(y) = \left(y - \frac{1}{N}\sum_{i=1}^{N}(y_i)\right)^2 = 1.29 \tag{31}$$

The relationship between MSE and R^2. can be shown in the following way. Look at the two equations: 32, 33:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2 \tag{32}$$

$$R^2 = 1 - \frac{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}{\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{33}$$

It's clear that the numerator of R^2 is equivalent to the expression of MSE and the denominator is equivalent to variance of y. Using this we get:

$$\frac{MSE}{Var(Y)} = \frac{0.012}{1.29} = 1 - R^2 = 1 - 0.01 \tag{34}$$

$$\Rightarrow R^2 = 0.99 \tag{35}$$

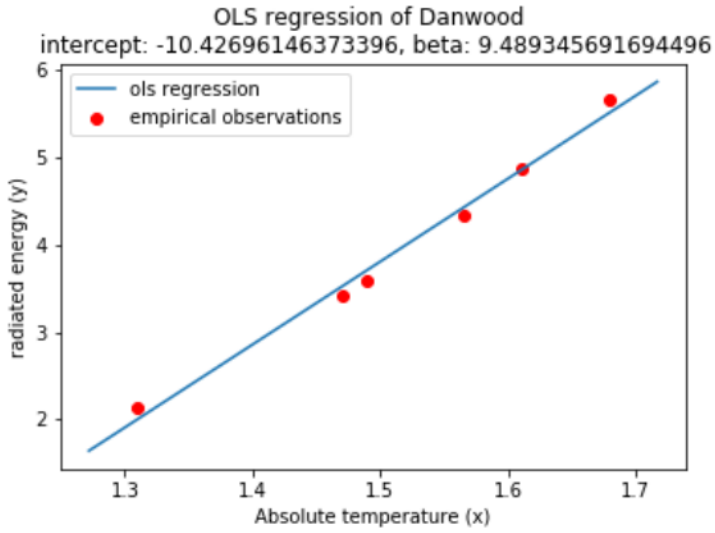
OLS regression of Danwood
intercept: -10.42696146373396, beta: 9.489345691694496
ols regression
empirical observations
radiated energy (y)
Absolute temperature (x)
6
5
4
3
2
1.3
1.4
1.5
1.6
1.7

Figure 4: