Danmarks Tekniske Universitet

(02450) Introduction to Machine Learning and Data Mining

Project 1

Silvia Rosvang Andersen (s214702)

Signe Djernis Olsen (s206759)

Sofie Kodal Larsen (s214699)

ID \ Section	1	2	3	4	5	6
s214702	20%	30%	40%	35%	40%	33.33%
s206759	30%	30%	40%	25%	40%	33.33%
s214699	50%	40%	20%	40%	20%	33.33%

Table 1: Responsibility for report

March 7th 2023

Contents

1 Introduction

The purpose of this project is to predict the weather in the city Albury in Australia. This report we will mainly focus on data exploration and principal component analysis presented by multiple visualizations. Through inspection, we decide to transform some of the data using the Box Cox transformation and thereafter standardize the data. A principal component analysis is then performed on the cleansed data.

From this data we want to apply machine learning and data mining methods to inspect which attributes influence on the weather prediction in Albury.

For further analysis we will apply classification and regression on the data. To do this we will make a regression model to predict if it will rain the next day depending on attributes in the data set. We will do a classification based on if it is going to rain the next day. From this we hope to be able to see a clear difference in some of the attributes and their influence on the rain prediction. But this will only be the aim for the next report.

2 Data set

The dataset consists of multiple weather attributes describing the weather from different cities in Australia spread out over 10 years. The data set consists of various attributes such as temperature, rainfall and windspeed. The data set was found on Kaggle.com [1] and the data is collected from the Bureau of Meteorology, Commonwealth of Australia.

The data set is widely used within the machine learning and data mining field, and many reports on prediction of the weather in Australia has been made. These reports focus on different methods to predict whether it is going to rain the next day, which will be out goal to predict as well. When looking at the results, the attributes that generally has the biggest impact on if it will rain the next day, is the humidity. One of the previous reports from the data was able to predict rainfall with a 93% accuracy.

2.1 Data cleansing

The raw data is a data frame containing 99516 rows (observations from all cities) and 23 columns (weather attributes). We choose to focus on the city Albury, so the data frame is reduced to 2142 rows.

The attributes RowID and Location doesn't contain any important information, so they are removed. The attributes Evaporation, Sunshine, Cloud9am and Cloud3pm consists mostly of NaN's and are therefore removed as well. We are not interested in the wind direction, since it is a nominal attribute with many factors, so WindGustDir, WindDir9am and WindDir3pm are also removed. The four attributes WindSpeed, Humidity, Pressure and Temperature have data from both 9am and 3pm, and we chose to reduce each attribute to the daily average.

Lastly we remove all the rows that contain NaN's. We also thought about substituting NaN' with the mean, but since only 22 out of 2142 rows contain NaN's we chose to delete them.

Now we have a new cleansed data frame `df` $\in \mathbb{R}^{2120\times 9}$.

[1] https://www.kaggle.com/datasets/arunavakrchakraborty/australia-weather-data

2.2 Explanation of the attributes

After data cleansing we have 9 attributes left we want to work with:

Variable	**Type of variable**	**Units**
MinTemp	Continuous Interval	Celsius
MaxTemp	Continuous Interval	Celsius
Rainfall	Continuous Ratio	Millimeters
WindGustSpeed	Continuous Ratio	Kilometers per hour
WindSpeed	Continuous Ratio	Kilometers per hour
Humidity	Continuous Ratio	Percent
Pressure	Continuous Ratio	Hectopascals
Temperature	Continuous Interval	Celsius
RainTomorrow	Binary Nominal	Yes(1)/no(0)

Table 2: Attributes after cleansing

To examine the data we look at the summary statistics. Table 3 shows an overview of each attribute and its belonging statistics. When looking at the table we can see that the median of Rainfall is 0 which indicates that there are more dry than wet days in Albury.

We know that MinTemp, MaxTemp and Temp are correlated since they all explain the weather condition on the same day. Therefore, when looking at the statistics we would expect the statistics of Temp to lie nicely between the statistics of MinTemp and Maxtemp, which they also do. Another attribute to look at is Pressure. The values of pressure are much higher than the values of the other attributes. To take this into account, we will standardize the data.

	MinTemp	MaxTemp	RainFall	WindGustSpeed	WindSpeed	Humidity	Pressure	Temp
Mean	9.47	22.52	2.01	32.84	11.28	61.20	1017.09	17.75
Std	6.04	7.78	6.54	13.42	5.93	17.39	7.11	6.74
Min	-2.10	7.50	0	11.00	0	18.00	986.35	4.85
25%	4.70	15.78	0	24.00	7.00	47.00	1012.29	11.85
50%	9.10	21.70	0	31.00	10.00	62.50	1017.03	17.25
75%	14.30	28.83	0.40	41.00	14.50	75.00	1021.85	23.00
Max	28.30	44.80	104.20	107.00	37.50	100.00	1037.95	37.15

Table 3: Summary of statistics

2.3 Standardization and data exploration

To work with our data, convert the data frame to a numpy array, $X \in \mathbb{R}^{2120\times 9}$, and the attribute names are stored in a separate array. We are now ready to explore the data. As mentioned, we will standardize the data so the influence of each attribute won't depend their numeric magnitude. Especially Pressure contains very high values compared to the other attributes. Therefor when doing PCA, Pressure would automatically have a high impact on the projection of the data, even though it might not be the most influential attribute.

The standardization has been carried out by subtracting the mean as follows

$$X^*_{j,i} = X_{j,i} - \mu_i \ \ , \ \ j \in \{1, .., 2120\}, \ i \in \{1, .., 9\}$$

To visualize the data, we make a box plot for all the attributes. Even though there are a lot of data points outside the 3rd quantile for Rainfall, these should not be considered as outliers.

Because of the standardization a lot of the observations in Rainfall will lie around 0, but the days it rained a lot will still stand out. Weather can be extreme, so it is normal that over several years, you have days with a lot of rain.

The same goes for the other attributes, so there are no clear outliers.

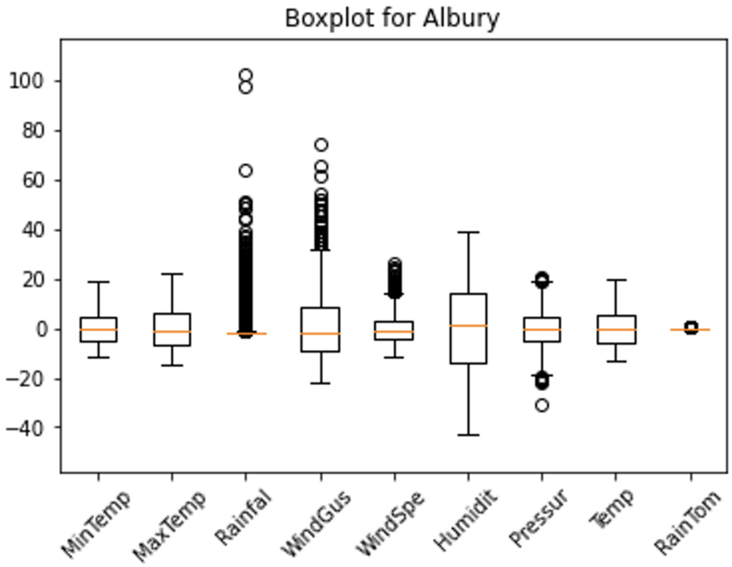


Figure 1: Boxplot for Albury

To get a clear visualization of distribution, the attributes are plotted as histograms i figure 2.

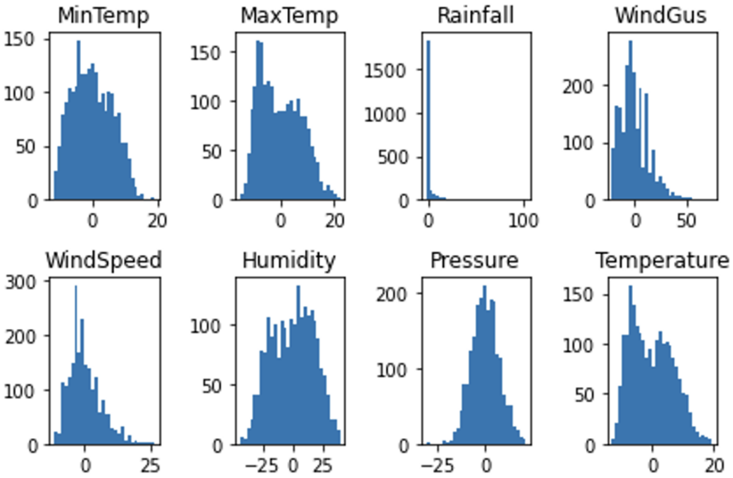


Figure 2: Histogram for all the attributes

From figure 1 and 2 we can see that MinTemp , Humidity and Pressure appears to be normally distributed. To investigate any possible correlation between attributes, we make a pairs plot shown in figure 3.

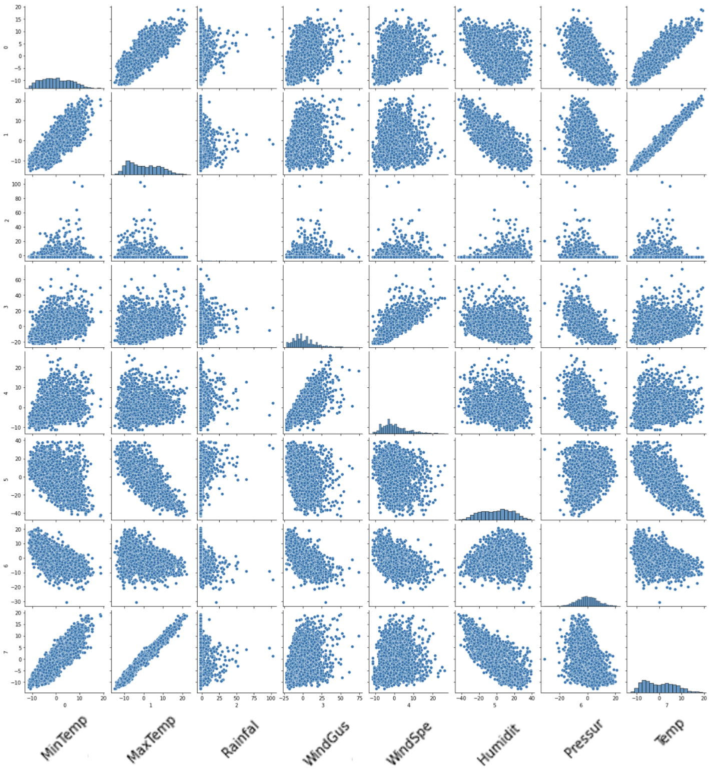


Figure 3: Correlations between the attributes

The pairs plot shows, that MinTemp and MaxTemp is strongly correlated, which makes great sense. It also seems, that Humidity is negatively correlated with Temperature, so that when one of the attributes have a high value the other will have a low value. Also WindGus and WindSpeed is naturally correlated. Rainfall does not seem to have a high correlation with the other attributes, but it's hard to tell, since it might need a transformation.

Some attributes, such as RainFall, WindGus and WindSpeed are not evenly distributed in the pairs plot, which might indicate, that they need to be transformed.

2.4 Transformation

If you look at the histograms in figure 2, it is clear that some attributes need to be transformed. The attributes WindGus and WindSpeed are both very right-skewed and needs a transformation. Since there are a lot of zero-elements in these attributes, we have decided to use the Box Cox transformation. First we shift the data to get only positive values and then apply the Box Cox transformation given by

$$f(y,\lambda) = \frac{y^\lambda - 1}{\lambda}.$$

Doing this we get the values

$$\lambda_{WindGust} = 0.432,\ \lambda_{WindSpeed} = 0.440.$$

After the transformation, the two attributes seem to fit a normal distribution. The histograms of the transformed and standardized data for WindGus and WindSpeed are shown in figure 4.

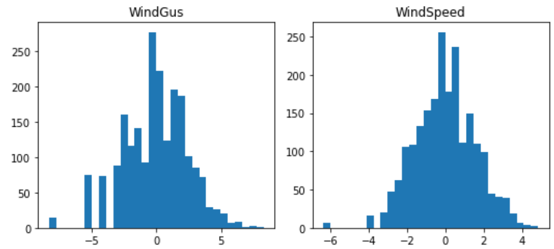


Figure 4: Histograms after transformation

If we look back at the histogram in figure 2, it seems that MaxTemp and Temperature follow a bimodal distribution, since they have two peaks. The reason for this could be, that summer and winter temperatures follow their own separate normal distribution. We also want to transform RainFall, but we haven't found a proper way of doing so, since the attribute contains a lot of zeros and a Box Cox transformation does not help in this case. Therefor, we have decided to leave it as it is.

In the following section we assume, the data is normally distributed.

3 PCA analysis

To perform PCA on our transformed and standardized data stored in the matrix $X^* \in \mathbb{R}^{2120\times 9}$ we compute singular value decomposition (SVD). The SVD is given by

$$X^* = U\Sigma V^T$$

where

- $U \in \mathbb{R}^{2120\times 2120}$
- $\Sigma \in \mathbb{R}^{2120\times 9}$
- $V \in \mathbb{R}^{9\times 9}$.

The matrix Σ is a diagonal matrix containing the singular values $\sigma_i, i = \{1..9\}$, for all the principal components (PC), and the matrix V contains all the vector projections of the PC's. We aren't going to use the matrix U, so we ignore this one. The SVD is computed and Σ for our data is given by

$$\Sigma = \begin{bmatrix} 0.7275 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.1469 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.0581 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.0501 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.0095 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.0056 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.0014 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.0001 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.0000 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

From the singular values we can calculate the amount of explained variance from the first K principal components with the formula

$$exp\ var. = \frac{\sum_{i=1}^{K} \sigma_i^2}{\sum_{i=1}^{M} \sigma_i^2}.$$

It turns out the first three principal components explain more than 90%, which is calculated by

$$\begin{aligned} exp.\ var &= \frac{\sum_{i=1}^{3} \sigma_i^2}{\sum_{i=1}^{9} \sigma_i^2} \\ &= \frac{0.7275 + 0.1469 + 0.0581}{0.7275 + 0.1469 + 0.0581 + 0.0501 + 0.0095 + 0.0056 + 0.0014 + 0.0001 + 0.0000} \\ &\approx 0.9325. \end{aligned}$$

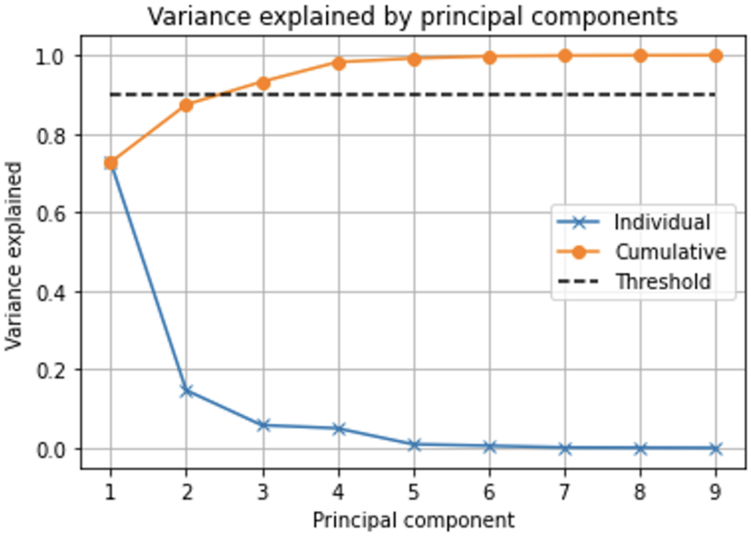


Figure 5: The variation explained by the PCA components

Figure 5 confirms that the first three PC's explain more than 90% of the variance, and the last four PC's explains barely nothing.

Since we only look at the first three PC's, we subtract the first three columns of V, contains the vector projections of the first three PC's, also called PCA component coefficients. The PCA component coefficients for PC one, two and three and given by

$$v_1 = \begin{bmatrix} -0.1991 \\ -0.3499 \\ 0.0714 \\ -0.0444 \\ -0.0197 \\ 0.8528 \\ 0.1173 \\ -0.2990 \\ 0.0065 \end{bmatrix}, v_2 = \begin{bmatrix} -0.4201 \\ -0.2056 \\ -0.4585 \\ -0.1302 \\ -0.0688 \\ -0.3231 \\ 0.6205 \\ -0.2439 \\ -0.01982 \end{bmatrix}, v_3 = \begin{bmatrix} 0.3007 \\ 0.3523 \\ -0.7387 \\ -0.0740 \\ -0.0695 \\ 0.3680 \\ 0.0612 \\ 0.3005 \\ 0.0066 \end{bmatrix}$$

Figure 6 visualizes the first three PCA component coefficients. It can be seen, that Humidity has the highest impact on the projecting of the data onto PC1, while the attributes MaxTemp, MinTemp, Pressure and Temp has roughly equal impact on the projection of PC3 numerically seen.

For PC2 and PC3 Rainfall has a high negative value, which means if the value of Rainfall is high and all the other attributes are low, we will have a negative value of the projection onto PC2 and PC3.

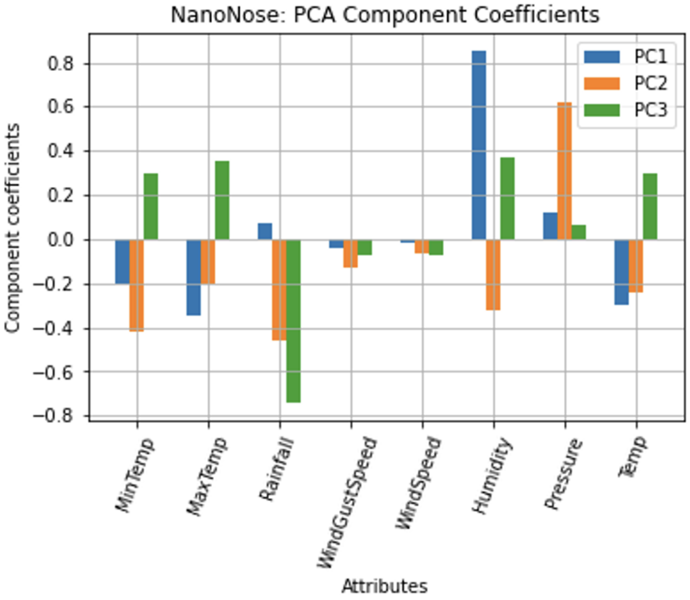
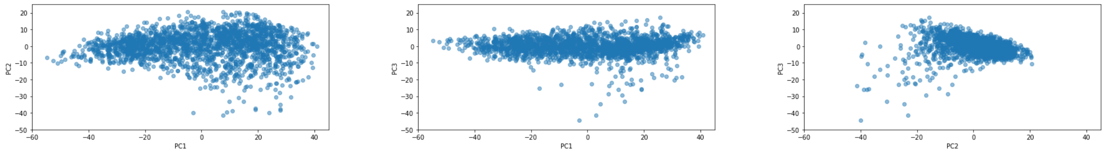


Figure 6: The principal directions of the PCA components

Overall, the PCA component coefficients tells us, that WindGustSpeed and WindSpeed have the lowest impact on the projection onto the three principal components, while Humidity and Rainfall have the highest.

3.1 Projected Data

In figure 7 we have visualized the data projected onto PC1, PC2 and PC3 in three 2D plots. We see that the projections of respectively on figure 7a and 7b the projected data is nicely spread, which means they explain a lot of the different features of the variance, especially 7a. If if we look at figure 7c the projection data is very clustered in the same area. This could correspond with the fact, that PC2 and PC3 explain less of the variance than PC1 does, and that PC2 and PC3 might explain some of the same variance.



(a) PC1 and PC2

(b) PC1 and PC3

(c) PC2 and PC3

Figure 7: Projected Data

4 Discussion

The raw data contains 23 attributes and 99516 rows. We decided to clean out the data to get a overview. Some choices had to be made, like choosing to look at one city (Albury), deleting some of the attributes and take the daily mean of the attributes with two observations per day. A minor part of the rows that contained NaN's were removed as well. All in all, we removed 14 attributes and were left with 9. The removed attributes could have contributed to the PCA and have given another result, which could be taken into account in further analysis.

The PCA assumes normal distribution for all attributes to get a good projection. When looking at our data it is clear that RainFall does not follow the normal distribution and also MaxTemp and Temperature would have needed a transformation. In the PCA we assumed normality for all the attributes. This false assumption of normal distribution might lead to some errors in the projections. Therefore it could be an idea to look into other ways to consider these attributes to make a more accurate projection.

It seems feasible to use this data for further analysis. The first three components from of PCA gives an explanation or more than 90% which indicates we can get a high explanation of variance from a low dimensional space. Therefore it will be possible to construct a regression model or make classification to predict whether it will rain the next day or not.

5 Conclusion

Some of the attributes didn't follow a normal distribution, which meant they had to be transformed. With the new transformed data, a principal component analysis could be computed. Here, only three components were needed to explain more than 90% of the variance of the data, so a reduction from a 9 to 3 dimensional vector space would only cause a loss of information about the data of less than 10%. A further reduction to a 1 dimensional space would have given an explanation of more than 70%.

For the three PCA first component coefficients, it could be seen that Rainfall and Humidity had some high numerical values, which means, they explain a lot of the variance in the data set. Other attributes, such as WindGustSpeed and WindSpeed had low component coefficients, and therefore don't explain a lot of the variance.

Through data cleansing and principal component analysis, it can be concluded, that the attributes MinTemp, MaxTemp, Pressure, Temp, and especially Rainfall and Humidity explain the most about the data.

6 Exam problems

6.1 Question 1

Option **D**.
Time of day, x_1, is meausred in a 30-minute interval and is an interval attribute. Number of traffic lights, x_6, and number of run over accidents, x_7, are both ratio, there is physical meaning of 0. Finally, the level of congestion, y, is an ordinal attribute. Therefor, the answer is option **D**.

6.2 Question 2

Option **A**.
In this question we need to predict the p-norm distance of two observations.
To do this we use the formula:

$$(|x_{14}(1) - x_{18}(1)|^p + |x_{14}(2) - x_{18}(2)|^p + ... + |x_{14}(7) - x_{18}(7)|^p)^{\frac{1}{p}}$$

with the special case of $p = \infty$

$$\max(|x_{14}(1) - x_{18}(1)|, |x_{14}(2) - x_{18}(2)|, ..., |x_{14}(7) - x_{18}(7)|)$$

A) Is $d_{p=\infty}(x_{14}, x_{18}) = 7.0$?
We calculate the $p = \infty$-norm

$$\max(|7|, |0|, |2|, |0|, |0|, |0|, |0|) = 7.$$

The answer **A** is hereby correct.

6.3 Question 3

Option **A**.
The explained variance is calculated by the first K out of all M principal components is given by the formula:

$$\textit{explained var.} = \frac{\sum_{i=1}^{K} \sigma_i^2}{\sum_{i=1}^{M} \sigma_i^2}.$$

The explained variance from the first 4 principal components is then given by

$$\textit{explained var.} = \frac{13.9^2 + 12.47^2 + 11.48^2 + 10.03^2}{13.9^2 + 12.47^2 + 11.48^2 + 10.03^2 + 9.45^2} = 0.87.$$

This means, the variance explained by the first four principal components is 0.87, which is greater than 0.8, and option **A** is true.

6.4 Question 4

Option **D**.
We want to determine which statements fit the data and PCAs.
We look at the projection onto principal component 2. For the data to have a positive projection onto PCA2, x_2, x_3, x_4 and x_5 should be high values and x_1 a low value.
Since this fits with the data, the correct answer is **D**.

6.5 Question 5

Option **A**.
We need to look at the Jaccard similarity of s_1 and s_2
The Jaccard similarity is defined as:

$$J(x,y) = \frac{f_{11}}{f_{11}+f_{01}+f_{10}}.$$

This means that the Jaccard similarity of s_1 and s_2 can be calculated as

$$J(s_1,s_2) = \frac{2}{2+6+5} = \frac{2}{13} = 0.153846.$$

or by using the total vocabulary size of M = 20000 which give the same

$$J(s_1,s_2) = \frac{f_{11}}{M-f_{00}} = \frac{2}{20000-(20000-13)} = \frac{2}{13} = 0.153846.$$

From the calculations we can conclude that the correct answer is **A**

6.6 Question 6

Option **B**.
The wanted the probability:

$$p(\hat{x}_2 = 0|y = 2).$$

So from looking at the given table, we see that $p(\hat{x}_2 = 0, \hat{x}_7 = 0|y = 2) = 0.81$ and $p(\hat{x}_2 = 0, \hat{x}_7 = 1|y = 2) = 0.03$
This means we can calculate $p(\hat{x}_2|y = 2)$ as follows

$$\begin{aligned} p(\hat{x}_2 = 0|y = 2) &= \frac{0.81+0.03}{0.81+0.03+0.1+0.06} \\ &= 0.84. \end{aligned}$$

The correct answer is option **B**.