# General notes

When calculating results using Octave, please provide code snippets by writing them on your worksheet. This is crucial for me to understand how you came up with your solution.

# 1 Conditional probabilities and Bayes' rule

**(15 pts)**

An approach to vowel classification is to determine the frequencies of the first two formants in the spectrogram ($f_1$ and $f_2$) and identify the best-matching vowel w.r.t. some training data. In a given training corpus, a number of sample recordings of the five vowels (a,e,i,o,u) have been analyzed and categorized into ranges of $f_1$ and $f_2$. The following table shows how many samples were associated with which range and which vowel in the corpus:

| ID | $f_{1,min}$[Hz] | $f_{1,max}$[Hz] | $f_{2,min}$[Hz] | $f_{2,max}$[Hz] | $n_a$ | $n_e$ | $n_i$ | $n_o$ | $n_u$ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 400 | 0 | 1200 | 0 | 3 | 1 | 5 | 7 |
| 2 | 0 | 400 | 1200 | 2000 | 8 | 7 | 6 | 5 | 0 |
| 3 | 0 | 400 | 2000 | $\infty$ | 1 | 6 | 7 | 0 | 5 |
| 4 | 400 | 800 | 0 | 1200 | 3 | 5 | 1 | 7 | 2 |
| 5 | 400 | 800 | 1200 | 2000 | 9 | 6 | 7 | 9 | 0 |
| 6 | 400 | 800 | 1200 | $\infty$ | 5 | 7 | 4 | 1 | 3 |
| 7 | 800 | $\infty$ | 0 | 1200 | 9 | 9 | 1 | 3 | 8 |
| 8 | 800 | $\infty$ | 1200 | 2000 | 9 | 8 | 7 | 0 | 5 |
| 9 | 800 | $\infty$ | 1200 | $\infty$ | 5 | 0 | 1 | 1 | 5 |

In a field test, the following formant sequence was observed:

$$
\begin{aligned}
f_1^8 &= \begin{pmatrix} f_{1,1} \\ f_{2,1} \end{pmatrix}, \dots, \begin{pmatrix} f_{1,8} \\ f_{2,8} \end{pmatrix} \\
&= \begin{pmatrix} 530 \\ 991 \end{pmatrix}, \begin{pmatrix} 1044 \\ 1960 \end{pmatrix}, \begin{pmatrix} 901 \\ 1977 \end{pmatrix}, \begin{pmatrix} 720 \\ 823 \end{pmatrix}, \begin{pmatrix} 235 \\ 728 \end{pmatrix}, \begin{pmatrix} 622 \\ 745 \end{pmatrix}, \begin{pmatrix} 371 \\ 590 \end{pmatrix}, \begin{pmatrix} 611 \\ 725 \end{pmatrix} [\text{Hz}]
\end{aligned}
$$

a) What is the most likely vowel sequence $v_{1(a)}^8$ associated with $f_1^8$?
Hint: First determine the ID sequence matching the formant sequence $f_1^8$.

The vowel classifier is supposed to be used as input interface for the 1986 IBM PC version of the game Tetris (as seen in the screenshot below). In order for this to work, the following association between vowels and actions is used:

| $v$ | action |
|---|---|
| e | l*e*ft |
| i | r*i*ght |
| a | rot*a*te |
| o | dr*o*p |


b) Considering the Tetris scenario, what is the most likely vowel sequence $v^8_{1(b)}$ associated with $f^8_1$?

From a vast number of games, statistics of game actions are available ($n$ is the number of tokens observed per action):

| $n/1000$ | action |
|---|---|
| 918 | left |
| 927 | right |
| 754 | rotate |
| 118 | drop |

c) Considering the Tetris scenario and the game action statistics, what is the most likely vowel sequence $v^8_{1(c)}$ associated with $f^8_1$?

## 2 Regression (15 pts)

The following table shows historic crude oil prices per barrel:

| year | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 |
|---|---|---|---|---|---|---|---|---|
| price [US$] | 28.5 | 24.4 | 25.0 | 28.8 | 38.2 | 54.5 | 65.1 | 72.3 |

Consider the regression functions (you should use $x$ = year − 2000 to avoid machine precision issues):

a) $y_a = w_0 + w_1 x$,

b) polynomial regression $y_b$ with $D = 2$,

c) polynomial regression $y_c$ with $D = 7$,

d) $y_d = w_0 + w_1 \sin\left(\frac{\pi}{4}x\right) + w_2 \sin\left(\frac{\pi}{4}(x+2)\right)$,

e) $y_e = w_0 + w_1 e^{\frac{x}{2}}$.

Given the above data of historic crude oil prices,

i) train linear regression models for each of the functions $y_a, \ldots, y_e$ and calculate the models' empirical risks $R_a, \ldots, R_e$.

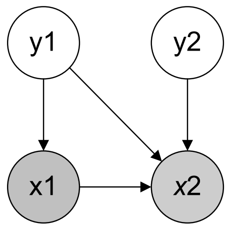
ii) Predict the average oil price for 2013 with each of the models.

iii) Which of the models fits the data best? Justify your answer.

iv) Which of the models fits the data worst? Justify your answer.

## 3 Bayesian networks (6 pts)

We are given the following Bayesian network:




Which of the following equations are correct w.r.t. this network? Which ones are wrong?

$$p(x_1, x_2) = p(x_1|y_1)p(x_2|y_1, y_2) \quad (1)$$

$$p(x_2) = \frac{p(x_1, x_2, y_1, y_2)}{p(x_1, y_1, y_2)} \quad (2)$$

$$p(y_2|y_1) = \frac{p(y_1|y_2)p(y_1)}{p(y_2)} \quad (3)$$

$$0 < \sum_{y \in Y_2} p(y_2 = y) < 1 \quad (4)$$

$$\max_{x_1} p(x_1|y_1) = \max_{x_1} p(y_1|x_1)p(x_1) \quad (5)$$

$$\arg\max_{x_1} p(x_1|y_1) = \arg\max_{x_1} c \cdot p(y_1|x_1)p(x_1) \quad \text{with} \quad c \in \mathbb{R} \quad (6)$$

Justify your answers. Hint: $Y_2$ is the set of all values the random variable $y_2$ can assume.

## 4 Naïve Bayes (20 pts)

A brand-new application to classification is to tell whether individuals are drunk or sober by analyzing speech data they are uttering. These are (Boolean) features extracted by the signal processing unit of an example implementation:

| $i$ | feature |
|---|---|
| 1 | high energy |
| 2 | high energy variance |
| 3 | high pitch |
| 4 | high pitch variance |
| 5 | high gitter |
| 6 | high shimmer |
| 7 | high speaking rate |
| 8 | affected by Lombard distortion |
| 9 | the word "please" is present |
| 10 | the word "f***" is present |

To train the classifier, $n(1)$ utterances have been manually marked as drunk $(d = 1)$ and $n(0)$ as sober $(d = 0)$. The following table shows how many of each feature were found to be true for both of the scenarios:

| $d$ | $n(d)$ | $n_1(d)$ | $n_2(d)$ | $n_3(d)$ | $n_4(d)$ | $n_5(d)$ | $n_6(d)$ | $n_7(d)$ | $n_8(d)$ | $n_9(d)$ | $n_{10}(d)$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 550 | 0 | 1 | 450 | 332 | 510 | 250 | 87 | 390 | 219 | 0 |
| 1 | 110 | 75 | 73 | 1 | 2 | 75 | 56 | 9 | 58 | 0 | 91 |

Build a naïve Bayes classifier.

a) Classify the following test samples:

$$\begin{aligned} x_1 &= (1, 1, 0, 0, 0, 0, 1, 1, 1, 1) \\ x_2 &= (1, 1, 1, 1, 1, 1, 0, 1, 1, 0) \\ x_3 &= (0, 1, 0, 1, 1, 1, 0, 1, 0, 1) \end{aligned}$$

In doing so, use a smoothing approach that replaces zero counts by a constant $\alpha = 1$.

b) Now, use a smoothing approach which increments every count by a constant $\beta = 1$. How do classification results change?