Solutions to exam 210316

1. (a) We know that the least-squares estimate $\widehat{\boldsymbol{\theta}}$ is the solution to the normal equations $\mathbf{X}^\mathsf{T}\mathbf{X}\widehat{\boldsymbol{\theta}} = \mathbf{X}^\mathsf{T}\mathbf{y}$. We can calculate these in Python as:

```
import numpy as np
X = np.array([[1,2],[3,4],[3,3],[1,1]])
y=[2, 8, 5, 3]

XTX=np.matmul(X.transpose(),X)
print(XTX)
XTy=np.matmul(X.transpose(),y)
print(XTy)

[[20 24]
 [24 30]]
[44 54]
```

From there we can see that the solution $\hat{\theta}_1 = 1$ and $\hat{\theta}_2 = 1$ satisfies they equations. For test data $x_1^* = 2$ and $x_2^* = 3$, we have $y^* = 2 + 3 = 5$.

(4p, 2 points for normal equations, 1 point for $\hat{\theta}$ and 1 point for test data)

(b) Further away. The regularization term makes $\|\widehat{\boldsymbol{\theta}}\|$ smaller, i.e., the estimated parameter values become closer to zero, and any change in $\|\widehat{\boldsymbol{\theta}}\|$ moves us away from the maximum likelihood estimate obtained by least squares.

(1p)

(c) Here any split which put points 1 and 4 (group L) and points 2 and 3 together together is minimal. The average for the two boxes is

$$\begin{aligned} \hat{L} &= (2+3)/2 = 2.5 \\ \hat{R} &= (5+8)/2 = 6.5 \end{aligned}$$

Depending on where the split was chosen, either 2.5 or 6.5 is the correct answer. (Average of the two for on boundary can also be considered correct).

(2p, 1 point for averages, 1 point for prediction)

(d) A model involving distance from centre squared, i.e. $x_1^2 + x_2^2$, is best, since it has one parameter and captures the relationship. $(x_1 + x_2)^2$ is also good, but less intuitive. Including both x_1^2 and x_2^2 is justifiable, but it should be explained why we have two variables. Distance, i.e. $\sqrt{x_1^2 + x_2^2}$, doesn't fully capture the relationship, where increase is squared. Including just x_1 and x_2 is wrong. (2p)

2. (a) We use

$$p(y \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid y)p(y)}{\sum_{m=1}^{M} p(\mathbf{x} \mid m)p(m)},$$

where $p(y)$ is the prior probability of class y, and $p(\mathbf{x} \mid y)$ the probability density of $\mathbf{x}$ for an observation from class y. (Bayes' theorem)

In our case $p(y = \text{Asthma}) = 0.1$. The probability density value for a score of 300 for someone with asthma is $\text{N}(300 \mid 81, 81) = 0.01274$ and for someone without asthma is $\text{N}(300 \mid 100, 100) = 0.05399$. Thus

$$p(y = \text{Asthma} \mid \mathbf{x}) = \frac{0.01274 \cdot 0.1}{0.01274 \cdot 0.1 + 0.05399 \cdot 0.9} \approx 0.0255$$

(4p)

(b) Negative scores are allowed for the Normal distribution but we can't blow a negative test. The standard deviation is so large that we do need to take this in to account.

(1p)

(c) If

$$p(X < x) = 1 - \exp(-x/\mu)$$

then the probability density function is

$$f(x|\mu) = \exp(-x/\mu)/\mu$$

Also using Bayes

$$\begin{aligned} p(\text{Asthma} \mid X = x) &= \frac{f(x|\mu = 81) \cdot 0.1}{f(x|\mu = 81) \cdot 0.1 + f(x|\mu = 100) \cdot 0.9} \\ &= \frac{\exp(-x/81) \cdot 0.1/81}{\exp(-x/81) \cdot 0.1/81 + \exp(-x/100) \cdot 0.9/100} \\ &= \frac{10}{10 + \exp(x/81 - x/100) \cdot 729/10} \end{aligned}$$

And finally,

$$p(\text{Asthma} \mid X = x) = \frac{1}{1 + \frac{9^3}{100} \exp\left(\frac{19x}{8100}\right)}$$

as required. (3p)

(d) $p(\text{Asthma} \mid X = 50) = 0.1087$ and $p(\text{Asthma} \mid X = 300) = 0.0635$. The probability that either of these people have asthma is not very different from the baseline at 0.1, so this test is not particularly useful (even if it is better than the Normal assumption).

(2p)

3. (a) We consider the two splits A and B in turn

Split A In the left branch we have 400 data points of class $Y = 0$ and 100 data points of class $Y = 1$. As a consequence we have in the right branch 200 data points of class $Y = 0$ and 500 data points of class $Y = 1$.

	Left branch	Right branch
Data points	$n_1 = 600$	$n_2 = 600$
Class $k = 0$	$\hat{p}_{10} = 500/600 = 5/6$	$\hat{p}_{20} = 100/600 = 1/6$
Class $k = 1$	$\hat{p}_{11} = 100/600 = 1/6$	$\hat{p}_{21} = 500/600 = 5/6$
Misclassification	$Q_1 = 1/6$	$Q_2 = 2/6$
Gini index	$Q_1 = 2 \cdot \frac{5}{36} = \frac{5}{18}$	$Q_2 = \frac{5}{18}$

The total cost with misclassification impurity measure is $n_1Q_1 + n_2Q_2 = 600/6 + 600/6 = 200$. The total cost with the Gini index impurity measure is $n_1Q_1 + n_2Q_2 = 600\frac{5}{18} + 600\frac{5}{18} = 6000/18$.

Split B In the left branch we have 600 data points of class $Y = 0$ and 200 data points of class $Y = 1$. As a consequence we have in the right branch 0 data points of class $Y = 0$ and 400 data points of class $Y = 1$.

	Left branch	Right branch
Data points	$n_1 = 600$	$n_2 = 400$
Class $k = 0$	$\hat{p}_{10} = 600/800 = 3/4$	$\hat{p}_{20} = 0/200 = 0$
Class $k = 1$	$\hat{p}_{11} = 200/800 = 1/4$	$\hat{p}_{21} = 600/600 = 1$
Misclassification	$Q_1 = 1/4$	$Q_2 = 0$
Gini index	$Q_1 = 6/16$	$Q_2 = 0$

The total cost with misclassification impurity measure is $n_1Q_1 + n_2Q_2 = 800/4 + 400 \cdot 0 = 200$. The total cost with the Gini index impurity measure is $n_1Q_1 + n_2Q_2 = 800 \cdot \frac{6}{16} = 300 < 6000/18$.

(4p, 2 points for the table, 2 points for the total cost and comparison.)

(b) Split B will most likely give the best conditions for further splitting the tree since one of the leaf nodes (the right) is already clean and does not need to be split further. If we use misclassification error both splits will have the same cost. If we use the Gini index we favour split B over split A.

(2p)

(c) See lecture notes for bagging description. It reduces variance but not bias.

(2p, 1 for description, 1 for variance reduction.)

(d) The 100 data points with $Y = 1$ on the left branch and the 100 data points with $Y = 0$ on the right branch

(2p, 1 point for left branch, 1 point for right branch.)

4. (a) The convolutional layer maps $5 \times 5 \times 3$ (filter rows $\times$ filter columns $\times$ input channels) to 24 output channels, meaning that $\mathbf{W}^{(1)} \in \mathbb{R}^{5\times5\times3\times24}$ and $\mathbf{b}^{(1)} \in \mathbb{R}^{24}$. Because of the stride 2, the hidden layer $\mathbf{q}$ will have dimension $10 \times 10 \times 24$.

   (2p, 2 points if all correct, order of the dimension should not cause any deduction of points)

   (b) The logits have dimension 4 (since there are 4 class probabilities to be predicted). The dense layer maps a vectorized version of $\mathbf{q}$ (dimension $10 \times 10 \times 24$) onto 4 logits $\mathbf{z}$, meaning $\mathbf{W}^{(2)} \in \mathbb{R}^{2400\times4}$ and $\mathbf{b}^{(2)} \in \mathbb{R}^4$.

   (2p, 2 points if all correct, order of the dimension should not cause any deduction of points)

   (c) In total, there are $5 \cdot 5 \cdot 3 \cdot 24 + 24 + 2\,400 \cdot 4 + 4 = 11\,428$ parameters.

   (d) In a dense layer each unit in the input layer is connected with *all* units in the output layer and each input-output-unit-pair has its *unique parameter*. In contrast, in a convolutional layer, each unit in the input layer is only connected with a *region of units* and all input units share the *same set of parameters*.

   (2p, 1 point for each of the properties (formulated in one way or another))

   (e) Since we have $10\,000$ data points and do a 90%-10% split we have $9\,000$ training data points and $1\,000$ validation data points. That means that we have a $100/9000 \approx 1.1\%$ misclassification rate on the training data but $100/1000 = 10\%$ miscalssification rate on validation data. The performance on validation data is substantially worse than the performance on training data. Hence we have overfitting problem. To improve the generalization on unseen data, i.e. improve the miscalssification rate on the validation data, we should try measures that help us deal with overfitting. These could be:

   - Reduce the number of filters in the convolutional layer. This makes the model less flexible.
   - Reduce the size of the filters in the convolutional layer to say 3×3. This makes the model less flexible.
   - Add regularization to the model.
   - Train for fewer number of epochs. For example, stop training when the validation error starts increasing (so called early stopping)
   - Collect more training data (if possible). This will improve the generalization to unseen data.

   (3p, 1 point for each suggested measure with correct motivation. Correct motivation but no suggested improvement that match this motivation scores 1p)

5. You are working for a company that works with building security and are developing a face recognition software.

   (a) The false positive rates are

   $$\frac{200}{800+200} = 0.2$$

   for both males and females. The true positive rates are

   $$\frac{90}{90+10} = 0.9 \text{ and } \frac{190}{190+10} = 0.95$$

   for males and females, respectively. And the precisions are

   $$\frac{90}{200+90} = 0.31 \text{ and } \frac{190}{190+200} = 0.49$$

   for males and females, respectively.

   (3p)

   (b) False positive rate is fair. True positive rate is unfair, probably against men who are locked out more often. Precision is possibly unfair against men, but it is difficult to see in what aspect it is unfair, so unless justified we should say it is 'neither'. Reasonable arguments in any direction are acceptable here.

   (2p)

   (c) Reading from the graph, precision is equal for both in the points furthest to the left, where precision (true positive rate) is 0.8. The other point (furthest to right) for where precision is equal has a false positive rate of 1, so shouldn't be used. Comparing to the AUC, we see that at this point the false positive rate is 0.1. It is hard to find a reasonable interpretation of precision in this particular application, and therefore (given the low true positive rate) it doesn't make sense to prioritise precision.

   (3p)

   (d) While the accuracy claims about true positive rate are correct, this does not mention the poor performance with regard to false positives. One in five people who should not be admitted to the building would be admitted erroneously, a point that would raise a lot of concerns for users. If we chose the method in the top left hand corner of the AUC chart then it would have a 90% accuracy on false positives and true positives for both males and females.

   (2p)