T01 - Statistics

1. 10 individuals

$H_0 : d_i = 0 \Rightarrow \mu_{before} = \mu_{after} \rightarrow \mu_0 = 0$
$H_1 : d_i \neq 0$

Paired t-test

$$t_0 = \frac{\bar{d} - \mu_0}{s/\sqrt{n}} \qquad \bar{d} = \frac{1}{n}\sum_i d_i = -8.1$$

$$s = \sqrt{\frac{\sum_i (d_i - \bar{d})^2}{n-1}} = 7.5931$$

$$t_0 = \frac{-8.1 - 0}{7.5931} \cdot \sqrt{10} \approx -3.3734$$

$\alpha = 0.05$

$t^c_{0.975;9} = 2.262 \longrightarrow |t_0| > t^c \Rightarrow$ H_0 is rejected suggesting that the tax increase affects the consumption.

2. 8 men and 10 woman $\Rightarrow$ independent $\rightarrow$ Welch's Test

$H_1 : \mu_f > \mu_m$ → on average, woman wear their mask longer than men
$H_0 : \mu_f \leq \mu_m$

$H_1 : \mu_D = \mu_f - \mu_m > 0$
$H_0 : \mu_D \leq 0$

- $t_0 = \dfrac{\bar{x_f} - \bar{x_m} - 0}{s_{\bar{x}_f - \bar{x}_m}}$

$$\bar{x_f} = \frac{4+2+3+5+\dots}{10} = 4$$

$$\bar{x_m} = \frac{2+1+5+\dots}{8} = 2.5$$

$$s_{\bar{x}_f - \bar{x}_m} = \sqrt{\frac{s_f^2}{n_f} + \frac{s_m^2}{n_m}}$$

$$s_f^2 = \frac{\sum (\bar{x}_f - x_{f,i})}{10 - 1} = 3.778$$

$$s_m^2 = 1.714$$

$$s_{\bar{f}-\bar{m}} = \sqrt{\frac{3.778}{10} + \frac{1.714}{8}} = 0.592$$

- $t_0 = \frac{4 - 2.5}{0.592} = 1.949$

$\rightarrow \alpha = 0.05 \quad df = 16 \quad \Rightarrow \quad t^c_{0.95;16} = 1.749$

$\therefore \; t_0 = 1.949 > 1.749 = t^c_{0.95;16}$

$\hookrightarrow H_0$ can be rejected $\rightarrow$ on average, women wear their mask longer.

3. a) H_0: The new technique does not improve the average test scores
H_1: The new technique improves the average test scores

$H_0: \mu_{NT} \leq \mu_{OT} \Leftrightarrow \mu_D \leq 0$
$H_1: \mu_{NT} > \mu_{OT} \Leftrightarrow \mu_D > 0$

b) This is one-sided because we are only interested in whether this new technique has a positive effect on the test scores.

c) $t = 8.8798$; $p\text{-value} = 6.1914 \times 10^{-10}$

$p\text{-value} < 0.05 \Rightarrow$ reject H_0

T02 - Regression

1. a) $\hat{\beta}_1 = \frac{cov(x,y)}{var(x)} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$

$= \frac{\sum_{i=1}^{n}(x_iy_i - \bar{y}x_i - \bar{x}y_i - \bar{y}\bar{x})}{\sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} - \bar{x}^2)} = \frac{(\frac{1}{n}\sum_{i=1}^{n} x_iy_i) - \bar{x}\bar{y}}{(\frac{1}{n}\sum_{i=1}^{n} x_i^2) - \bar{x}^2}$

$= \frac{\frac{1}{9}\times 2.914.3 - 3.78 \cdot 63}{\frac{1}{9}\times 262.22 - 3.78^2} \approx 5.77$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 x \approx 63 - 5.77 \cdot \bar{x} \approx 41.19$

$\therefore \hat{y}(x) = 5.77\,x + 41.19$

b) β_0 : A country with a capital gross national product of 0 has (approximately) 41.19% of literate people

β_1 : With the increase of \$1.000, the percentage of literate people grows (approximately) 5.77%

c) $H_0 : \beta_1 \leq 0$ $\quad H_1 : \beta_1 > 0$ $\quad \rightarrow$ one-sided

$t_0 = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$ $\quad SE(\hat{\beta}_1) = \sqrt{\frac{RSS}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \cdot \frac{1}{n-2}} \approx 2.17$

$= \frac{5.77}{2.17} \approx 2.66$ $\quad RSS = \sum_{i=1}^{n}(\hat{y}_i - y_i) = 4411.4$

$\sum (x_i - \bar{x})^2 = 133.77$

$\alpha = 0.05$, $\quad df = n-2 = 7$

$t^c_{0.95;7} = 1.895 \quad \Rightarrow \quad t_0 > t^c_{0.95;7}$

$\therefore$ We can reject H_0 and conclude that β_1 is statistically significant.

d) A prediction for countries outside the sample range might lead to percentages of literate people that do not make sense (i.e. $<0\%$ and $>100\%$)

T03 - Logistic Regression

1. a)

$$
\begin{aligned}
& p(y=1|x) = p(y=0|x) \\
\Leftrightarrow\ & p(y=1|x) = 1 - p(y=1|x) \\
\Leftrightarrow\ & p(y=1|x) = 1/2 \\
\Leftrightarrow\ & \sigma(\beta_0 + \beta_1 x_1 + \beta_2 x_2) = 1/2 \\
\Leftrightarrow\ & \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}} = 1/2 \\
\Leftrightarrow\ & 1/2 + 1/2\, e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2)} = 1 \\
\Leftrightarrow\ & e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2} = 1 \\
\Leftrightarrow\ & \beta_0 + \beta_1 x_1 + \beta_2 x_2 = 0
\end{aligned}
$$

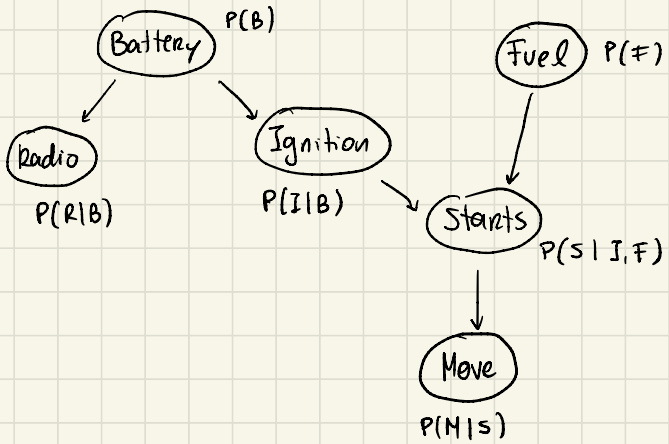
We have $x_2 = f(x_1)$

$$
\Leftrightarrow \quad x_2 = -\frac{\beta_0 + \beta_1 x_1}{\beta_2} = -\frac{\beta_0}{\beta_2} - \frac{\beta_1}{\beta_2} x_1
$$

b) 6 miss-classified samples

TD4 - Naïve Bayes

1. a)



b) $P(B, F, I, M, R, S) = P(M|S) \cdot P(S|I,F) \cdot P(I|B) \cdot P(F) \cdot P(R|B) \cdot P(B)$

2. a) age: continuous
preferences: categorical / discrete
money: continuous
product: discrete

	continuous / discrete	range	scale of measurement
age	continuous	$[0, +\infty[$	ratio
preferences	discrete	$\{R, B, C\}$	nominal
money	continuous	$[0, +\infty[$	ratio
product	discrete	$\{UC, VC, NP\}$	nominal

b) Estimate prior → count instances

$P_{UC} = 0.622$ $P_{VC} = 0.208$ $P_{NP} = 0.170$

c) age:

λ_{UC} = data.loc[data["product"] == "Unicorn", "age"].mean()
= 14.9147

$\lambda_{UC} = 56.0817$

$\lambda_{NP} = 29.8$

preferences:

$\Pi_{UC} =$ data.loc [data["product"] == "Unicorn", "preferences"].
value_counts (normalize = True)

↳ $\Pi^R_{UC} = 0.579$; $\Pi^R_{VC} = 0.356$; $\Pi^R_{NP} = 0.441$

...

money:

$\eta_{UC} = 0.092$ $\quad\quad$ $\eta = \bar{x}^{-1}$

$\eta_{VC} = 0.066$

$\eta_{NP} = 0.109$

d) $P(\text{product} \mid \text{age}, \text{preferences}, \text{money}) =$

$$= \frac{P(\text{age} \mid \text{product}) \cdot P(\text{preferences} \mid \text{product}) \cdot f(\text{money} \mid \text{product}) \cdot P(\text{product})}{Z}$$

Anna : age = 81 ; money = 53.10 € ; preferences = "Cats"

↳ plug in the values for each class/product and compute the probabilities

e) 4th customer → only know that they like rainbows

$$P(\text{product} \mid \text{preferences}) = \frac{P(\text{preferences} == \text{"Rainbows"} \mid \text{product}) \cdot P(\text{product})}{Z}$$

- plug product == "UC"

TO5 - Decision Trees

1. a) $\text{entropy}((0.1, 0.9)) = -0.1 \log_2(0.1) - 0.9 \log_2(0.9) = 0.469$

b) $\text{entropy}((0.8, 0.2)) = -0.8 \log_2(0.8) - 0.2 \log_2(0.2) = 0.722$

c) $\text{entropy}((0.5, 0.5)) = -2 \cdot 0.5 \log_2(0.5) = 1.0$

e) $\text{info}([2,3]) = \text{entropy}((0.4, 0.6)) = 0.971$

h) $\text{info}([2,3],[9,0]) = \frac{5}{14}\text{info}([2,3]) + \frac{9}{14}\text{info}([9,0])$

$= \frac{5}{14}\text{entropy}\left(\left(\frac{2}{2+3}, \frac{3}{2+3}\right)\right) + \frac{9}{14}\overbrace{\text{entropy}((1,0))}^{\text{pure} \rightarrow \text{entropy} = 0}$

2. a) possible good splits:

$s \in (35,37) \rightarrow \text{info}([0,2],[3,1]) = \frac{4}{6}\text{entropy}((0.75, 0.25)) = 0.541$ ↙

$s \in (37,40) \rightarrow \text{info}([1,2],[2,1]) = 2 \cdot \text{entropy}((0.66, 0.33)) = 0.918$

→ Optimal split: $s = 36$

3.

a) $S_1 = \{x \mid \text{Past\_Trend}(x) = \text{Positive}\}$

$S_2 = \{x \mid \text{Past\_Trend}(x) = \text{Negative}\}$

$\frac{|S_1|}{|S|} = \frac{3}{5} \qquad \frac{|S_2|}{|S|} = \frac{2}{5}$

$\frac{\{x \in S_1 \mid \text{Return}(x) = \text{Up}\}}{|S_1|} = \frac{2}{3} \qquad \frac{\{x \in S_1 \mid \text{Return}(x) = \text{Down}\}}{|S_1|} = \frac{1}{3}$

$G(S_1) = 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 = \frac{4}{9}$

$G(S_2) = 1 - (1)^2 - (0)^2 = 0$

$G(S_1, S_2) = \frac{3}{5} \times \frac{4}{9} + \frac{2}{5} \times 0 = \frac{4}{15}$

Open Interest :

High → $G(S_1) = 1 - 0.5^2 - 0.5^2 = 1/2$

Low → $G(S_2) = 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2 = 1 - \frac{1}{9} - \frac{4}{9} = \frac{4}{9}$

$$G(S_1, S_2) = \frac{4}{10} \times \frac{1}{2} + \frac{6}{10} \times \frac{4}{9} = \frac{2}{10} + \frac{12}{45} = \frac{3}{15} + \frac{4}{15} = \frac{7}{15}$$

b) Choose the one with the lowest Gini Index →

T07 - Model Evaluation and Selection

1.

a)

Actual \ Predicted	0	1
0	3	2
1	3	2

b) $\text{Recall} = \frac{TP}{TP+FN} = \frac{2}{5}$

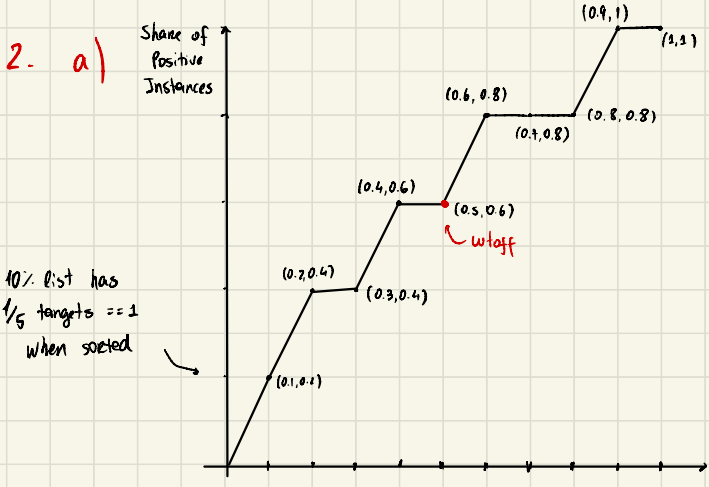
$\text{Precision} = \frac{TP}{TP+FP} = 0.5$

$\text{Specificity} = \frac{TN}{TN+FP} = 0.6$

$\text{False Alarm Rate} = \frac{FP}{FP+TN} = 0.4$

$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} = 0.5$

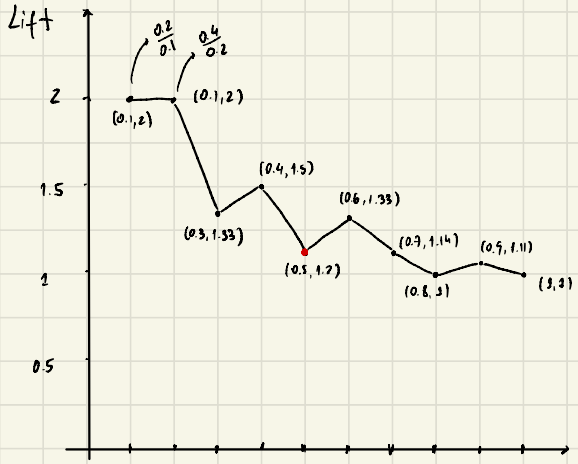
2. a)



- It's a poor wtoff choice since a higher value (>0.65) would capture the same amount of positive instances with a smaller share of the data

Gain Curve

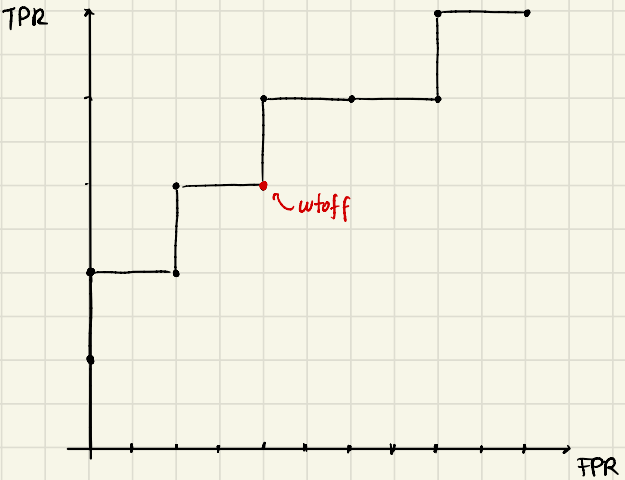
b)



- Cutoff at (0.5, 1.2) is a poor choice because it is a local minimum.

Lift Curve

c)



- Cutoff corresponds to the point (0.4, 0.6). It is a poor choice, as a lower cutoff would increase the TPR and keep the same FPR.

ROC Curve

d)

- Colleague evaluates the model on subset of training set which may cause overfitting since the model might not generalize well for unseen data.
- The lift curve will not be a constant. It will start at 2.5, and then monotonically decrease with the first "-" instance.
- My lift curve is not monotonically decreasing. It is neither monotonically increasing nor decreasing.

4. a) 50%.

b) Leave-one-out → Positive instance
32-1 positive 32 negative ⇒ majority = Negative } 100% error rate

Leave-one-out → Negative instance
32-1 negative 32 positive ⇒ majority = Positive } 100% error rate

∴ Averaging over all 64 folds, error remains 100%.

c) We drastically overestimate the error of the majority classifier. This example showcases a possible failure of leave-one-out cross validation.

TO8 - Clustering

1. 1. Calculate cluster distance from every point to each cluster

$x_1 = (2.5, 3)$

$d_{1A} = \sqrt{(2.5-0.5)^2 + (3-2.5)^2}$
$= \sqrt{2^2 + 0.5^2} = 2.061$

$d_{1B} = \sqrt{2.5^2 + 0.5^2} = 2.549$

$d_{1C} = \sqrt{2^2 + 2.5^2} = 3.201$

Closest cluster: A
assigned

For each point we have:

$x_2 = B$
$x_3 = A$
$x_4 = A$
$x_5 = B$
$x_6 = A$
$x_7 = A$
$x_8 = C$
$x_9 = C$
$x_{10} = A$
$x_{11} = A$
$x_{12} = C$

2. Update cluster centers using the assigned items

$$c_A = \frac{\begin{bmatrix}2.5\\3\end{bmatrix} + \begin{bmatrix}2\\2.75\end{bmatrix} + \begin{bmatrix}2.5\\2.5\end{bmatrix} + \begin{bmatrix}0.5\\1\end{bmatrix} + \begin{bmatrix}1\\1\end{bmatrix} + \begin{bmatrix}0.75\\0.75\end{bmatrix} + \begin{bmatrix}1\\0.5\end{bmatrix}}{7}$$

$$= \begin{bmatrix}1.464\\1.643\end{bmatrix}$$

$$c_B = \begin{bmatrix}3.0\\2.75\end{bmatrix} \qquad c_C = \begin{bmatrix}3.42\\0.83\end{bmatrix}$$

3. Reassign points to the near cluster

4. Recalculate cluster centroids

5. Reassign points ⇒ No changes, STOP!

2. a) $x_0 = (1,1)$ $\quad d_{0,A} = \sqrt{9+0} = 3$ $\quad d_{0,B} = \sqrt{9+1} > 3$

↳ assign: A

$x_1 = (1,2)$ $\quad d_{1,A} = \sqrt{9+1} > 0$ $\quad d_{1,B} = \sqrt{9+0} = 3$

↳ assign: B

$x_2 = (7,1)$ $\quad d_{2,A} = \sqrt{9+0} = 3$ $\quad d_{2,B} = \sqrt{9+1} > 0$

↳ assign: A

$x_3 = (7,2)$ $\quad d_{3,A} = \sqrt{9+1} > 0$ $\quad d_{3,B} = \sqrt{9+0} = 3$

↳ assign: B

new centroids: $c_A = \frac{\begin{bmatrix}1\\1\end{bmatrix} + \begin{bmatrix}7\\1\end{bmatrix}}{2} = \begin{bmatrix}4\\1\end{bmatrix}$ $\quad c_B = \frac{\begin{bmatrix}1\\2\end{bmatrix} + \begin{bmatrix}7\\2\end{bmatrix}}{2} = \begin{bmatrix}4\\2\end{bmatrix}$

Next reassignment will be equal ⇒ terminate!

Based on the results, we can see that given the initial centroids, the result won't be the "obvious" clusters that would group (0,1) and (2,3) together.

↳ Solution: Multiple iterations with different random initial centroids.

b) K-means will end after third iteration with (3,20) being assigned to A and the centroid recalculated. The first cluster now coincides with the outlier and is not representative of its observations. In particular, K-means is sensitive to outliers and dependent of the number of the chosen number of clusters.

↳ Solution: multiple iterations with different number K of clusters or alternative clustering approach

3.

	(0,1,4,6)	(2,7)	(5,8,9)	(3)
(0,1,4,6)	0	3	4	6
(2,7)		0	7	11
(5,8,9)			0	6
(3)				0

Smallest distance is $d((0,1,4,6),(2,7))$ so we merge those 2 clusters.

	(0,1,2,4,6,7)	(5,8,9)	(3)
(0,1,2,4,6,7)	0	$\min(4,7)=4$	6
(5,8,9)		0	6
(3)			0

Smallest dist is $d((0,1,2,4,6,7),(5,8,9)) \Rightarrow$ merge

...

Smallest dist (and only possible) $d((0,1,2,4,5,6,7,8,9),(3)) \Rightarrow$ merge

∴ Hierarchical clustering complete!

T09 - Principal Component Analysis

1. $D = \begin{bmatrix} -3 & -1 & -1 \\ 0 & -1 & 0 \\ -2 & -1 & 2 \\ 1 & -1 & 3 \end{bmatrix}$

a) $\bar{\mu}_1 = \frac{-3+0-2+1}{4} = -1 \qquad \bar{\mu}_2 = -1 \qquad \bar{\mu}_3 = 1$

$X = \begin{bmatrix} -2 & 0 & -2 \\ 1 & 0 & -1 \\ -1 & 0 & 1 \\ 2 & 0 & 2 \end{bmatrix} = D - \bar{\mu}$

b) $\text{Var}(x_j) = \frac{1}{n-1} \sum_{i=1}^{N} (x_{ij} - \bar{x}_j)^2 \qquad \text{cov}(x_j, x_k) = \frac{1}{n-1} \sum_{i=1}^{N} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)$

$\text{Var}(x_1) = \frac{1}{3} \times (2^2 + 1^2 + 1^2 + 2^2) = 10/3$

→ Note: The mean now is 0 since we transformed the dataset to a zero means.

$\text{cov}(x_1, x_2) = \frac{1}{3} \times 0 = 0$

$\text{cov}(x_1, x_3) = \frac{1}{3}(4 - 1 - 1 + 4) = 2$

$$\Sigma = \begin{bmatrix} \text{Var}(x_1) & \text{cov}(x_1, x_2) & \text{cov}(x_1, x_3) \\ \text{cov}(x_2, x_2) & \text{Var}(x_2) & \text{cov}(x_2, x_3) \\ \text{cov}(x_3, x_1) & \text{cov}(x_3, x_2) & \text{Var}(x_3) \end{bmatrix} = \begin{bmatrix} 10/3 & 0 & 2 \\ 0 & 0 & 0 \\ 2 & 0 & 10/3 \end{bmatrix}$$

c) $|\Sigma_X - \lambda I_3| = 0 \Leftrightarrow \det\left(\begin{bmatrix} 10/3 - \lambda & 0 & 2 \\ 0 & -\lambda & 0 \\ 2 & 0 & 10/3 - \lambda \end{bmatrix} \right) = 0$

$\Leftrightarrow \left(\frac{10}{3} - \lambda\right)(-\lambda)\left(\frac{10}{3} - \lambda\right) + (0 \cdot 0 \cdot 2) + (2 \cdot 0 \cdot 0) - (2 \cdot 2 \cdot (-\lambda)) - (0 \cdot 0 \cdot (\frac{10}{3} - \lambda)) - ((\frac{10}{3} - \lambda) \cdot 0 \cdot 0) = 0$

$$\Leftrightarrow \left(\lambda^2 - \frac{10}{3}\lambda\right)\left(\frac{10}{3} - \lambda\right) + 4\lambda \Leftrightarrow -\lambda^3 + \frac{10}{3}\lambda^2 - \frac{64}{9}\lambda + \frac{10}{3}\lambda^2 = 0$$

$$\Leftrightarrow \lambda\left(-\lambda^2 + \frac{20}{3}\lambda - \frac{100}{9}\right) = 0 \Leftrightarrow \lambda = 0 \quad \vee \quad -9\lambda^2 + 60\lambda - 64 = 0$$

$$\Leftrightarrow \lambda = 0 \quad \vee \quad \lambda = \frac{-60 \pm \sqrt{60^2 - 4\cdot(-9)\cdot(-64}}{2\cdot(-9)} = 0$$

$$\Leftrightarrow \lambda_1 = 0 \quad \vee \quad \lambda_2 = \frac{4}{3} \quad \vee \quad \lambda_3 = \frac{16}{3}$$

d) $(\Sigma_x - \lambda_3 I_3)v = 0 \Leftrightarrow \Sigma_x v = \lambda_3 v$

$\lambda = \frac{16}{3}$

$$\begin{bmatrix} -2 & 0 & 2 \\ 0 & -\frac{16}{3} & 0 \\ 2 & 0 & -2 \end{bmatrix} \cdot \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = 0 \Leftrightarrow \begin{cases} -2v_1 + 2\cdot v_3 = 0 \\ -\frac{16}{3} v_2 = 0 \\ 2v_1 - 2v_3 = 0 \end{cases} \Leftrightarrow \begin{cases} v_1 = v_3 \\ v_2 = 0 \end{cases}$$

Possible $v = k\begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$, $k \in \mathbb{R}$ $\Rightarrow$ Normalize $v = \begin{bmatrix} \sqrt{2}/2 \\ 0 \\ \sqrt{2}/2 \end{bmatrix}$

$\lambda = 0$

$$\begin{bmatrix} \frac{10}{3} & 0 & 2 \\ 0 & 0 & 0 \\ 2 & 0 & \frac{10}{3} \end{bmatrix} \cdot v = 0 \Leftrightarrow \begin{cases} \frac{10}{3} v_1 + 2v_3 = 0 \\ 2v_2 + \frac{10}{3} v_3 = 0 \end{cases} = \begin{cases} v_1 = v_3 = 0 \\ v_2 \text{ any value} \end{cases}$$

Possible vector $v = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$ ← Normalized

$\lambda = \frac{4}{3}$

$$\begin{bmatrix} 2 & 0 & 2 \\ 0 & -\frac{4}{3} & 0 \\ 2 & 0 & 2 \end{bmatrix} \cdot v = 0 \Leftrightarrow \begin{cases} 2v_1 + 2v_3 = 0 \\ -\frac{4}{3} v_2 = 0 \\ 2v_1 + 2v_3 = 0 \end{cases} \Leftrightarrow \begin{cases} v_1 = -v_3 \\ v_2 = 0 \end{cases}$$

Possible vector: $v = \begin{bmatrix} \frac{\sqrt{2}}{2} \\ 0 \\ -\frac{\sqrt{2}}{2} \end{bmatrix}$

- The eigenvectors are ordered in increasing order by corresponding eigenvalues

$$\Phi = \begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} & 0 \\ 0 & 0 & 1 \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} & 0 \end{bmatrix}$$

The proportions of variance are

- λ_3 explains $\frac{\lambda_3}{\lambda_1+\lambda_2+\lambda_3} = 80\%$ of the variance
- λ_2 explains 20% of the variance
- λ_1 explains 0% $\Rightarrow$ just λ_3 and λ_2 explain all the variance

e) $Z = X \cdot \Phi$ 1D

$$= \begin{bmatrix} -2 & 0 & -2 \\ 1 & 0 & -1 \\ -1 & 0 & 1 \\ 2 & 0 & 2 \end{bmatrix} \cdot \begin{bmatrix} \frac{\sqrt{2}}{2} \\ 0 \\ \frac{\sqrt{2}}{2} \end{bmatrix} = \begin{bmatrix} -2\sqrt{2} \\ 0 \\ 0 \\ 2\sqrt{2} \end{bmatrix}$$

f) 2D $\begin{bmatrix} -2 & 0 & -2 \\ 1 & 0 & -1 \\ -1 & 0 & 1 \\ 2 & 0 & 2 \end{bmatrix} \cdot \begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ 0 & 0 \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{bmatrix} = \begin{bmatrix} -2\sqrt{2} & 0 \\ 0 & \sqrt{2} \\ 0 & -\sqrt{2} \\ 2\sqrt{2} & 0 \end{bmatrix}$

2. a) $D \approx Z \cdot \Phi^T +$ means

$$\approx \begin{bmatrix} -2\sqrt{2} \\ 0 \\ 0 \\ 2\sqrt{2} \end{bmatrix} \cdot \begin{bmatrix} \frac{\sqrt{2}}{2} & 0 & \frac{\sqrt{2}}{2} \end{bmatrix} + \begin{bmatrix} -1 & -1 & 1 \\ -1 & -1 & 1 \\ -1 & -1 & 1 \\ -1 & -1 & 1 \end{bmatrix} = \begin{bmatrix} -3 & -1 & -1 \\ -1 & -1 & 1 \\ -1 & -1 & 1 \\ 1 & -1 & 3 \end{bmatrix}$$

- We would lose some information after this reconstruction since we considered only one dimension, omitting the second one when projecting

b) $D \approx Z \cdot \Phi^T + \text{means}$ where Φ contains v_3 and v_2

- Since $v_2 + v_3$ explain all the variance, the reconstruction leads to no information loss. ($D \approx Z \cdot \Phi^T + \text{means} = D$)

3. a) PCA : zero-mean X, eigenvalues, eigenvectors, Φ ordered by λ_i desc.

b) $D \approx Z \cdot \Phi^T + \text{means}$ with $Z = X \cdot \Phi$

c) add $(a, 5)$ so that it can be fully explained by the first pc. Any projected point from b) would work

d) 1. changes both eigenvalues and eigenvectors. Scaling only one point would change the zero-mean and covariances matrix.

2. $K^2 \cdot \Sigma_X \cdot v = \underbrace{K^2 \cdot \lambda}_{\text{scaled eigenvalues}} \cdot v$ but eigenvectors wouldn't change since they would be normalized anyways.

3. "Flipping" coordinates will result on reflection of the p.c

ex: $v_1 = \begin{bmatrix} 0.55 \\ -0.83 \end{bmatrix} \Rightarrow v_1' = \begin{bmatrix} -0.83 \\ 0.55 \end{bmatrix}$

T11 - Convex Optimization

1. $f(x, y) = a \exp(3x) + \frac{b}{2} xy + y^2$

$$\nabla f = \begin{pmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{pmatrix} = \begin{pmatrix} 3a \exp(3x) + \frac{b}{2} y \\ \frac{b}{2} x + 2y \end{pmatrix}$$

$$\nabla^2 f = \begin{pmatrix} \frac{\partial^2 f}{\partial^2 x} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial^2 y} \end{pmatrix} = \begin{pmatrix} 9a\exp(3x) & \frac{b}{2} \\ \frac{b}{2} & 2 \end{pmatrix}$$

f is convex iff $\nabla^2 f(x,y)$ is positive semidifinite. This is the case if all the principal minors are non-negative.

- First minor $H_1(x,y) = 9a\exp(3x) \Rightarrow H_1 \geq 0 \Leftrightarrow a \geq 0$
- $H_2(x,y) = 2 \geq 0$
- $H_3(x,y) = 18a\exp(3x) - \frac{b^2}{4} \geq 0$ for $a \geq 0$ and $b = 0$

$18a\exp(3x) \geq \frac{b^2}{4}$ → there might be an x that does not sustain this

$\therefore$ f is convex iff $a \geq 0$ and $b = 0$

2. convex iff $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$

- $$\begin{aligned} h(\lambda x + (1-\lambda)y) &= g_1(A(\lambda x + (1-\lambda)y) + b) \\ &= g_1(A\lambda x + (1-\lambda)Ay + b) \\ &= g_1(\lambda Ax + Ay - \lambda Ay + b + \lambda b - \lambda b) \\ &= g_1(\lambda(Ax + b) + (1-\lambda)(Ay + b)) \\ &\leq \lambda g_1(Ax + b) + (1-\lambda) g_1(Ay + b) \\ &= \lambda h(x) + (1-\lambda) h(y) \end{aligned}$$

3. $f(x,y) = 2x^2 + 0.5y^2 - 3x - y - 2xy + 5$ $\qquad z^{(1)} = (0,0)$

$$\nabla f(x,y) = \begin{pmatrix} 4x - 3 - 2y \\ y - 1 - 2x \end{pmatrix}$$

- $\nabla f(0,0) = \begin{pmatrix} -3 \\ -1 \end{pmatrix}$ $\qquad$ line search: $\alpha = \operatorname{argmin} f(z^{(1)} - \alpha \nabla f(z^{(1)}))$

$$= \operatorname{argmin} f(3\alpha, \alpha)$$

$$= \operatorname{argmin} 12.5\alpha^2 - 10\alpha + 5 \quad \cup$$

$f(3\alpha, \alpha) = 18\alpha^2 + 0.5\alpha^2 - 9\alpha - \alpha - 6\alpha^2 + 5$ ↗ $\qquad \hookrightarrow \alpha^* = 0.4$

$\nabla f(3\alpha, \alpha) = 25\alpha - 10 = 0 \Rightarrow \alpha^* = \frac{10}{25} = 0.4$

$z^{(2)} = z^{(1)} - \alpha^* \nabla f(z^{(1)}) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} - 0.4 \cdot \begin{pmatrix} -3 \\ -1 \end{pmatrix} = \begin{pmatrix} 1.2 \\ 0.4 \end{pmatrix}$

Step 2: $z^{(2)} = \begin{pmatrix} 1.2 \\ 0.4 \end{pmatrix}$ $\quad f(z^{(2)}) = 3$

$\nabla f(z^{(2)}) = \begin{pmatrix} 1 \\ -3 \end{pmatrix}$

$\alpha^* = \text{argmin } f(z^{(2)} - \alpha \nabla f(z^{(2)}))$
$= \text{arg min } f(1.2 - \alpha, 0.4 + 3\alpha)$
$= \text{arg min } 12.5\alpha^2 - 10\alpha + 3 \quad \Rightarrow \alpha^* = 0.4$

$\nabla f(1.2 - \alpha, 0.4 + 3\alpha) = 25\alpha - 10 = 0 \Rightarrow \alpha^* = 0.4$

$z^{(3)} = z^{(2)} - \alpha \cdot \nabla f(z^{(2)}) = \begin{pmatrix} 1.2 \\ 0.4 \end{pmatrix} - 0.4 \cdot \begin{pmatrix} 1 \\ -3 \end{pmatrix} = \begin{pmatrix} 0.8 \\ 1.6 \end{pmatrix}$

T12 - SGD and Neural Networks

1. 1.

- $K = 1 \Rightarrow F(x) = W_1 x$ → linear with no bias TRUE
- $K = 2$ $\quad \nabla \ell(F) = \frac{\partial \ell}{\partial F} \cdot \frac{\partial F}{\partial x} = -\frac{2}{n} \sum_{i=1}^{n} (y_i - F(x_i)) \cdot (-W_1 W_2)$
$= (W_1 W_2) \cdot \frac{2}{n} \sum_{i=1}^{n} (y_i - F(x_i))$
TRUE
- Increasing $K \Rightarrow F(x) = W_K W_{K-1} \dots W_1 x$
↳ all linear transformations can be mapped to $W_1 = W_K W_{K-1} \dots W_1$
TRUE
- $W_K \in \mathbb{R}^{d_1 \times d_2}$ $(d_1, d_2 > 1)$ leads to output of size $d_2 \times K$ for a K that depends on the previous results size

$F(x) \in \mathbb{R}^{1 \times 1}$ since it outputs a scalar label
$\therefore d_1 > 1$ but needs to be 1 FALSE

2. $\frac{\partial l}{\partial w_3} = \frac{\partial l}{\partial F} \cdot \frac{\partial F}{\partial w_3} = -2(y - F(x)) \cdot w_1 w_2 x$

2. a) $N = (2 \times 2 + 2) + (2 \times 1 + 1) = 9$ trainable parameters

b) $\hat{y} = \sigma(\sigma(X \cdot W^{[2]} + \vec{1} \cdot b^{[1]}) W^{[2]} + \vec{1} \times b^{[2]})$

c) $z^{[1]} = X \cdot W^{[1]} + \vec{1} \times b^{[1]}$

$$= \begin{pmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \end{pmatrix}$$

$$A^{[1]} = g_1(z^{[1]}) = \sigma(z^{[1]}) = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.731 \\ 0.731 & 0.5 \\ 0.731 & 0.731 \end{pmatrix}$$

$$z^{[2]} = A^{[1]} \cdot W^{[2]} + b^{[2]} = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.731 \\ 0.731 & 0.5 \\ 0.731 & 0.731 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ -1 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ -0.231 \\ 0.231 \\ 0 \end{pmatrix}$$

$$\hat{y} = A^{[2]} = \sigma(z^{[2]}) = \begin{pmatrix} 0.5 \\ 0.443 \\ 0.557 \\ 0.5 \end{pmatrix}$$

$$\mathcal{L}(y, \hat{y}) = \frac{1}{n} \sum_{i=1}^{n} l(y_i, \hat{y}_i)$$

$$= \frac{1}{4}\left(-\left[(1-0) \cdot \ln(1-0.5)\right] - \left[1 \cdot \ln(0.443)\right] - \left[1 \cdot \ln(0.557)\right] - \left[1 \cdot \ln(0.5)\right]\right)$$

$$\approx 0.696$$

d) $W^{[1]}, W^{[2]}, b^{[1]}, b^{[2]}$

$$\frac{\partial L}{\partial W^{[1]}} = \frac{\partial L}{\partial \hat{Y}} \cdot \frac{\partial \hat{Y}}{\partial Z^{[2]}} \cdot \frac{\partial Z^{[2]}}{\partial W^{[2]}} \qquad \frac{\partial L}{\partial b^{[2]}} = \frac{\partial L}{\partial \hat{Y}} \cdot \frac{\partial \hat{Y}}{\partial Z^{[2]}} \cdot \frac{\partial Z^{[2]}}{\partial b^{[2]}}$$

$$\frac{\partial L}{\partial W^{[1]}} = \frac{\partial L}{\partial \hat{Y}} \cdot \frac{\partial \hat{Y}}{\partial Z^{[2]}} \cdot \frac{\partial Z^{[2]}}{\partial A^{[1]}} \cdot \frac{\partial A^{[1]}}{\partial Z^{[1]}} \cdot \frac{\partial Z^{[1]}}{\partial W^{[1]}}$$

$$\frac{\partial L}{\partial b^{[1]}} = \frac{\partial L}{\partial \hat{Y}} \cdot \frac{\partial \hat{Y}}{\partial Z^{[2]}} \cdot \frac{\partial Z^{[2]}}{\partial A^{[1]}} \cdot \frac{\partial A^{[1]}}{\partial Z^{[1]}} \cdot \frac{\partial Z^{[1]}}{\partial b^{[1]}}$$

$$\frac{\partial L}{\partial \hat{Y}_j} \cdot \frac{\partial \hat{Y}_j}{\partial Z_j^{[2]}} \quad (\sigma) = -\left[\frac{Y_j}{\hat{Y}_j} - \frac{1-Y_j}{1-\hat{Y}_j}\right] \cdot \left[\hat{Y}_j(1-\hat{Y}_j)\right] = \hat{Y}_j - Y_j$$

$$\frac{\partial L}{\partial A^{[2]}} \cdot \frac{\partial A^{[2]}}{\partial Z^{[2]}} \quad (\hat{Y}) = \frac{1}{4} \cdot \vec{1}^{T} \cdot \operatorname{diag}\left[\left(A_j^{[2]} - Y_j\right)_{j \in \{1,2,3,4\}}\right]$$

$$= (0.125 \quad -0.139 \quad -0.111 \quad -0.125) \in \mathbb{R}^{1\times 4}$$

We have $\frac{\partial Z^{[i]}}{\partial W^{[i]}} = A^{[i-1]}$ and $\frac{\partial Z^{[i]}}{\partial b^{[i]}} = 1$

So $\frac{\partial L}{\partial W^{[2]}} = (0.125 \quad -0.139 \quad -0.111 \quad -0.125) \cdot \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.731 \\ 0.731 & 0.5 \\ 0.731 & 0.731 \end{pmatrix}$

$$= (-0.179 \quad -0.186) \in \mathbb{R}^{1\times 2}$$

$$\frac{\partial L}{\partial b^{[2]}} = (0.125 \quad -0.139 \quad -0.111 \quad -0.125) \cdot \vec{1} = -0.25 \in \mathbb{R}^{1\times 1}$$

...

e)

$$W^{[2]}_{new} = W^{[2]} - \alpha \cdot \left(\frac{\partial L}{\partial W^{[2]}}\right)^{T} = \begin{pmatrix} 1 \\ -1 \end{pmatrix} - 1.0 \begin{pmatrix} -0.179 \\ -0.186 \end{pmatrix} = \begin{pmatrix} 1.179 \\ -0.814 \end{pmatrix}$$

$$b^{[2]}_{new} = b^{[1]} - \alpha \cdot \left(\frac{\partial L}{\partial b^{[2]}}\right)^{T} = (0) - 1.0\,(-0.25) = (0.25)$$

$$W^{[1]}_{new} = W^{[1]} - \alpha \cdot \left(\frac{\partial L}{\partial W^{[1]}}\right)^T$$

$$= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - 1.0 \begin{pmatrix} -0.046 & 0.052 \\ -0.059 & 0.052 \end{pmatrix} = \begin{pmatrix} 1.046 & -0.052 \\ 0.059 & 0.948 \end{pmatrix}$$

d) $\hat{y}_{new} = g^{[2]}\left(g^{[1]}\left(X \cdot W^{[1]}_{new} + \vec{1} \cdot b^{[1]}_{new}\right) \cdot W^{[2]}_{new} + \vec{1} \cdot b^{[2]}_{new}\right)$

$$L(y, \hat{y}) = \ldots = 0.585$$

As expected, the empirical risk is lower than before. If the learning rate α is not sufficiently small, however, it can happen that the empirical risk increases after a weight update.

Homework 1

1. $H_0: \mu_X = 9.5$
   $H_1: \mu_X \neq 9.5$

   one sample, known variance → z-test

$$z_0 = \frac{\bar{x} - \mu}{\sigma_0/\sqrt{n}} = \frac{10.5 - 9.5}{2.5/\sqrt{25}} = \frac{1}{2.5} \times 5 = 2$$

$\alpha = 0.05$ and $\alpha = 0.01$, two-sided

$z^c_{0.975} = 1.96 \rightarrow z_0 > 1.96 \rightarrow H_0$ is rejected

$z^c_{0.995} = 2.58 \rightarrow z_0 < 2.58 \rightarrow H_0$ is <u>not</u> rejected

2. two groups: non-carnivore, carnivore
   known σ and $\bar{x}$, normally dist

a) Group 1: $\left[\bar{x}_1 \pm t_{0.975;11} \cdot \frac{s}{\sqrt{n}}\right] = \left[1780 \pm 2.201 \cdot \frac{236}{\sqrt{12}}\right] = [1633.86, 1926.14]$

Group 2: $\left[\bar{x}_2 \pm t_{0.975;19} \cdot \frac{s}{\sqrt{n}}\right] = [1783.00, 2017.00]$

b) The confidence intervals overlap to a great extend, making a decisive inference about the hypothesis not possible.

c) $H_0: \mu_1 < \mu_2$ ⇒ non-carnivore intake is lower than carnivore
$H_1: \mu_1 \geq \mu_2$

Known μ's, unknown σ's, 2 groups ↳ Welch's test

$H_0: \mu_d < 0 : \mu_1 - \mu_2 < 0$
$H_1: \mu_1 \geq 0$

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}} = 24.88 \approx 25$$

With $\alpha = 0.05$ and a one-tail test, we get $t^c_{0.95,25} = -1.708$

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{-120}{86.79} \approx -1.38$$

Since we have $t_0 > t^c$, we cannot reject H_0. This means, regarding the correctness of H_1, no conclusion can be drawn from the hypothesis test.

3. 1) single sample, Known mean with unknown σ

2) $H_0: \mu_x = \mu_0 = 0$

3) t-test: $\bar{x} = 2.65$ $s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \approx 2.55$

$$t_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{2.65}{\sqrt{2.55}} \cdot \sqrt{20} \approx 7.42148$$

4) $t^c_{0.975;19} = 2.093$

5) Since $t_0 > t^c$, we can reject the null hypothesis (H_0).

Homework 2

1. a) $\hat{\beta}_1 = \frac{cov(t, Demand)}{var(t)} = \frac{91.38}{7.5} = 12.1833$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 74.1155 - 12.1833 \cdot 4 = 25.3823$

$\hat{y}_{10} = \hat{\beta}_0 + \hat{\beta}_1 \cdot 10 = 147,2152$

b) $RMSE = \sqrt{MSE} = \sqrt{\frac{RSS}{T}} \simeq 1.755$

with $RSS = \sum_{i=1}^{n} (y - \hat{y})^2 = e_t^2$

Homework 3

2. a) $\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x = -3.89 + 0.135 \cdot x$

$e^{0.135} = \frac{odds(age+1)}{odds(age)} = odds\ ratio = 1.145$

$\Leftrightarrow \frac{p}{1-p} = e^{\beta_0 + \beta_1 x} \Rightarrow e^{\beta_0 + \beta_2 (x+1)} = e^{\beta + \beta_1 x} \times e^{\beta_1}$ ✓ growth

∴ When age increases by +1, the odds of having a car increase by 14.5%

b) $\ln\left(\frac{p}{1-p}\right) = \ln\left(\frac{p'}{1-p'}\right) \Leftrightarrow p(1-p') = p'(1-p) \Rightarrow p = p'$

when $p = 0.5 \Rightarrow \ln\left(\frac{0.5}{0.5}\right) = \ln(1) = 0$

$\beta_1 x + \beta_0 = 0 \Leftrightarrow -3.89 + 0.135x = 0 \Leftrightarrow x = 28.8$

Age = 28.8 is the number where the model is indifferent between the two classes

Homework 4

1. a) $\pi_{rent} = \frac{3}{8}$ $\pi_{property} = 1 - \frac{3}{8} = \frac{5}{8}$

b) $P(J=e \mid LS=R) = \frac{2}{3}$ $P(J=f \mid LS=R) = \frac{1}{3}$

$P(J=e \mid LS=P) = \frac{2}{5}$ $P(J=f \mid LS=P) = \frac{3}{5}$

...

c)

$$P(LS=R \mid J=e, MS=S, C=yes) = \frac{P(J=e \mid LS=R) \cdot P(MS=S \mid LS=R) \cdot P(C=yes \mid LS=R) \cdot P(LS=R)}{Z}$$

$$P(LS=P \mid J=e, MS=S, C=yes) = \frac{P(J=e \mid LS=P) \cdot P(MS=S \mid LS=P) \cdot P(C=yes \mid LS=P) \cdot P(LS=P)}{Z}$$

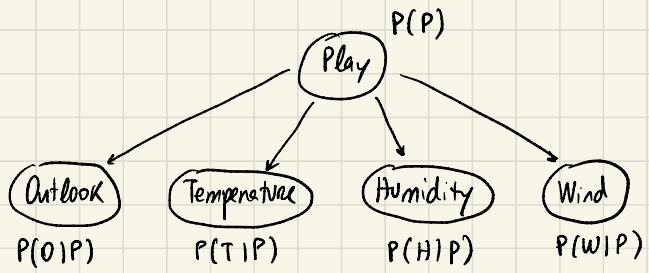
Since $P(LS=R \mid \dots) = \frac{1}{Z} \cdot \frac{1}{36} < P(LS=P \mid \dots) = \frac{1}{Z} \cdot \frac{3}{50}$ the classifier would classify it as a property owner.

2. a) $P(P=no \mid sunny, mild, normal, \overline{wind}) = \frac{1}{x} \cdot \frac{3}{5} \cdot \frac{2}{5} \cdot \frac{1}{5} \cdot \frac{2}{5} \cdot \frac{5}{14} = \frac{1}{x} \cdot \frac{60}{8750}$

$P(P=yes \mid sunny, mild, normal, \overline{wind}) = \frac{1}{x} \cdot \frac{2}{9} \cdot \frac{4}{9} \cdot \frac{6}{9} \cdot \frac{3}{9} \cdot \frac{9}{14} = \frac{1}{x} \cdot \frac{2592}{91854}$

↳ $P(P=no \mid \dots) \gg P(P=yes \mid \dots)$ so #15 is classified as No.

b)



Homework 5

1. $gain\_ratio = \frac{gain(x)}{info(x)}$ $\quad IG(Result) = info([7,3,4]) = 1.493$

- $gainRatio(Host) = \frac{gain(Host)}{intrinsic\_info(Host)} = \frac{info([7,3,4]) - info([5,2,1],[2,1,3])}{info([8,6])}$

$$= \frac{entropy(\frac{7}{14},\frac{3}{14},\frac{4}{14}) - \left(\frac{8}{14}\, entropy(\frac{3}{8},\frac{2}{8},\frac{1}{8}) + \frac{6}{14}\, entropy(\frac{2}{6},\frac{1}{6},\frac{3}{6})\right)}{entropy(\frac{8}{14},\frac{6}{14})}$$

$$entropy(\tfrac{7}{14},\tfrac{3}{14},\tfrac{4}{14}) = -\tfrac{1}{2}\log(\tfrac{1}{2}) - \tfrac{3}{14}\log(\tfrac{3}{14}) - \tfrac{2}{7}\log(\tfrac{2}{7})$$

. . .

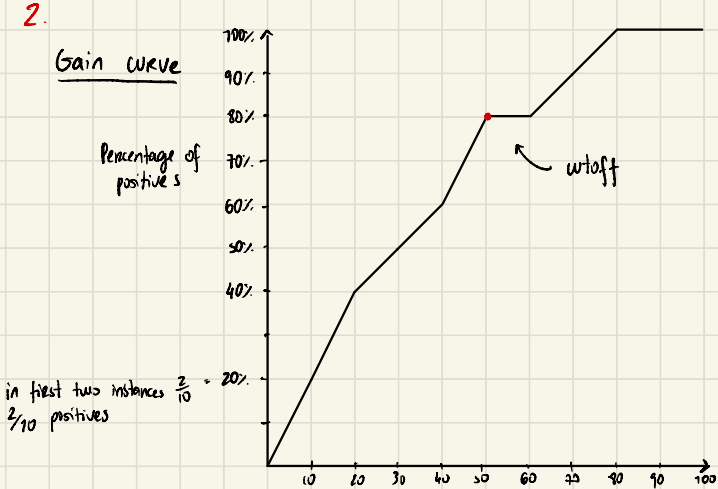
→ first split (Tradition ≤ 2.5) since it yields the greatest IG

Homework 7

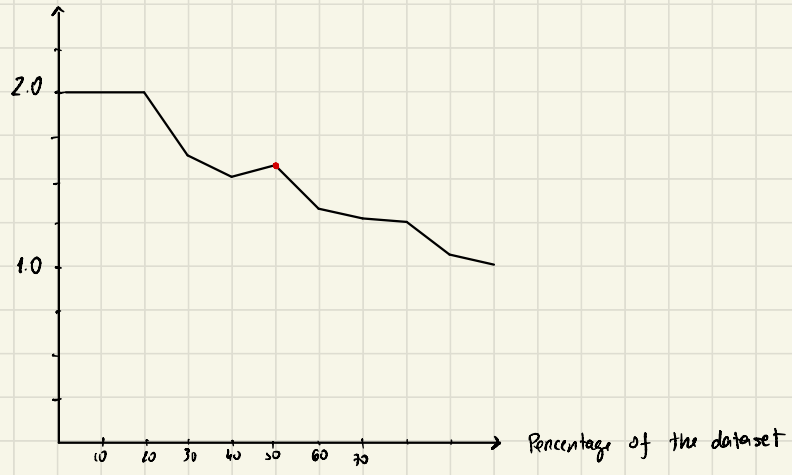
1. $Accuracy = \frac{TP+TN}{\#All} = \frac{10}{15} = \frac{2}{3}$ $\quad FPR = False\ Alarm\ Rate = \frac{FP}{FP+TN} = \frac{2}{7}$

$TPR = Recall = \frac{TP}{TP+FN} = \frac{5}{8}$ $\quad TNR = Specificity = \frac{TN}{TN+FP} = \frac{5}{7}$

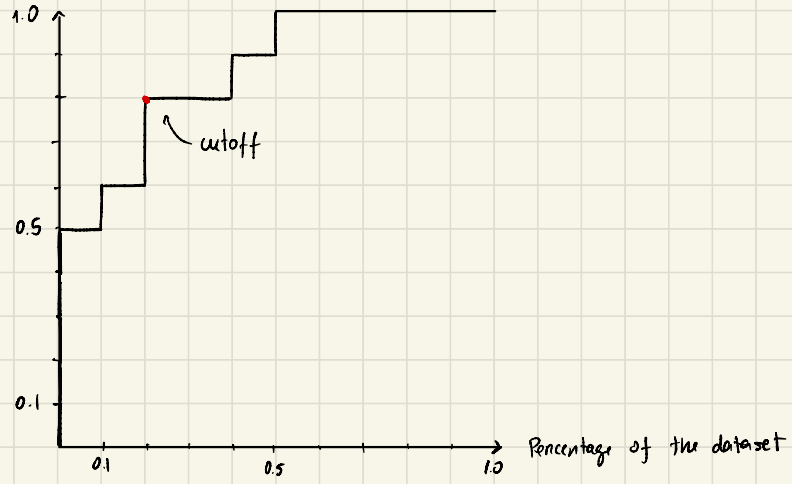
2.



Lift curve



ROC



3. Stratified 5-fold cross-validation
   - Class-distribution : [10, 10]
   - 20 entries
   - 5 folds

} folds with 4 items that keep 1:1 label distribution

Step	Training	Testing
1	P2 ∪ P3 ∪ P4 ∪ P5	P1
2	P1 ∪ P3 ∪ P4 ∪ P5	P2
3	P1 ∪ P2 ∪ P4 ∪ P5	P3
4	P1 ∪ P2 ∪ P3 ∪ P5	P4
5	P1 ∪ P2 ∪ P3 ∪ P4	P5

Homework 8

1. a) L1 norm : $\| x_i - x_j \|_1$

$$d_1(0,1) = |x_0 - x_1| + |y_0 - y_1| = 1 + 5 = 6$$

$$d_1(1,6) = |1 - 1| + |5 - (-5)| = 0 + 10 = 10$$

b) L2 norm : $\| x_i - x_j \|_2 = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2}$

$$d_2(0,9) = \sqrt{(0-(-1))^2 + (0-(-1))^2} = \sqrt{2} = 1.414$$

c) bottom-up hierarchical clustering

- join 0 with 7 → {0,7}
- row with {0,7} will have in each entry : $\min \{ d_1(0,i), d_1(7,i) \}$
- $\min (D_1^{(1)}) = 3$
  - cluster (0,7) with 9
  - 1 with 3 / or 5
  - 2 with 5

d) Rule of thumb: distance between the clusters is large. Graphically speaking, we draw a horizontal line in dendogram such that the longest lines are cut.

2. 1. Calculate probabilities for all points

$$P(A|x) = \frac{P(x|A) \cdot P(A)}{P(x)} \qquad P(B|x) = \frac{P(x|B) \cdot P(B)}{P(x)}$$

$$P(x|A) = f(x_1 ; \mu_A, \sigma_A) = \frac{1}{\sigma_A \sqrt{2\pi}} e^{-\frac{(x_1 - \mu_A)^2}{2\sigma_A^2}} = 0.374$$

$$P(x|B) = f(x_1 ; \mu_B, \sigma_B) = 0.2015$$

$$P(x) = P(x|A) \cdot P(A) + P(x|B) + P(B) = 0.2015$$

2. Update parameters:

$$\mu_A = \frac{\sum_{i=1}^{6} P(A|x_i)\, x_i}{\sum_{i=1}^{6} P(A|x_i)} = 1.10$$

$$\sigma_A = \frac{\sum_{i=1}^{6} P(A|x_i)\,(x_i - \mu_A)^2}{\sum_{i=1}^{6} P(A|x_i)} = 0.44$$

$$p_A = \frac{\sum_{i=1}^{6} P(A|x_i)}{6} = 49\%$$

3. Recalculate cluster probabilities
4. Update distribution parameters
5. Calculate cluster probabilities ← probabilities!

⇒ No change in cluster assignment → termination!

3. a) combined "expert" models → more stable, less variance
↳ better prediction performance

b) bagging: draws samples from dataset and train different models in parallel
boosting: improve last models predictions by giving more importance to misclassified samples by the previous model
→ not parallelizable, generally more overfitting

Homework 9

1. a) $f(\lambda) = \lambda^3 - 8.5\lambda^2 + 15\lambda$ eigenvalues are the roots of $f(\lambda)$

$$f(\lambda) = \lambda(\lambda^2 - 8.5\lambda + 15) = 0$$

$$\Leftrightarrow \lambda = 0 \quad \vee \quad \lambda = \frac{+8.5 \pm \sqrt{8.5^2 - 4 \cdot 15}}{2}$$

$$\hookrightarrow \lambda_1 = 0 \quad \vee \quad \lambda_2 = 2.5 \quad \vee \quad \lambda_3 = 6$$

λ_1 : 0% variance $\qquad$ λ_3 : 70.59%

$\lambda_2 : \frac{2.5}{6+2.5} = 29.41\%$ variance

→ only 2 components needed to explain all the variance

b) $\lambda_3 = 6$

$$(\Sigma_x - \lambda I_3)\, v = 0 \Leftrightarrow \begin{bmatrix} -1 & 2 & -1 \\ 2 & -4.7 & 0.6 \\ -1 & 0.6 & -3.8 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = 0 \Leftrightarrow \begin{cases} -v_1 + 2v_2 - v_3 = 0 \\ 2v_1 - 4.7\, v_2 + 0.6 v_3 = 0 \\ -v_1 + 0.6\, v_2 - 3.8 v_3 = 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} v_1 = 2v_2 - v_3 \\ v_1 = 2.35\, v_2 - 0.3 v_3 \\ \text{—} \end{cases} \Leftrightarrow \begin{cases} 0.35 v_2 = -1.7 v_3 \\ \text{—} \\ \text{—} \end{cases}$$

...

$$v = R \begin{bmatrix} 5 \\ 2 \\ -1 \end{bmatrix} \quad \text{where} \quad R \in \mathbb{R}$$

→ Normalization : $\begin{bmatrix} 0.913 \\ 0.365 \\ -0.183 \end{bmatrix}$

Zero-mean dataset : $\bar{x_1} = 1$, $\bar{x_2} = 1.4$, $\bar{x_3} = 1.8$ ⇒ obtain X

projection $Z = X \cdot \Phi$ where $\Phi = V$ (eigenvector for $\lambda = 6$)

c) multicollinearity, violates Gauss-Markov properties. One feature is a linear combination of the other 2.

d) mean changes → cov changes → $p(\lambda)$ change → eigenvalues change

$$P = \begin{bmatrix} 4 - 1 \\ 2.6 - 1.4 \\ 1.2 - 1.8 \end{bmatrix} = \begin{bmatrix} 3 \\ 1.2 \\ -0.6 \end{bmatrix} = 3.286\, V$$

→ Exactly along the calculated PC. ⇒ eigenvectors / pc won't change

Homework 11

1. $f(x,y) = \exp(ax+by^2)$

$$\nabla f(x,y) = \begin{pmatrix} a\exp(ax+by^2) \\ 2b\exp(ax+by^2) \end{pmatrix} \qquad \nabla^2 f(x,y) = \begin{pmatrix} a^2\exp(ax+by^2) & 2aby\exp(ax+by^2) \\ 2aby\exp(ax+by^2) & (4b^2y^2+2b)\exp(ax+by^2) \end{pmatrix}$$

First principal minor: $a^2\exp(ax+by^2) \geq 0$, $\forall x,y \in \mathbb{R}$

Second: $(4b^2y^2+2b)\exp(ax+by^2) \geq 0 \Leftrightarrow 4b^2y^2+2b \geq 0 \quad \forall x,y \in \mathbb{R}$ if $b \geq 0$

Third (det): $H_3 \geq 0$ for all $x,y \in \mathbb{R}$ iff $b \geq 0$

$\rightarrow \nabla^2 f(x,y)$ positive semidefinite

$\therefore f(x,y)$ is convex $\forall a \in \mathbb{R}, b \geq 0$

2. b) $f(x,y) = \frac{1}{2}x^2 + \exp(-y) + 3xy$

$$\nabla f(x,y) = \begin{pmatrix} x+3y \\ -\exp(-y)+3x \end{pmatrix} \qquad \nabla^2 f(x,y) = \begin{pmatrix} 1 & 3 \\ 3 & \exp(-y) \end{pmatrix}$$

$H_1: 1 \geq 0$; $H_2 = \exp(-y) \geq 0$; $H_3 = \exp(-y) - 9 \not\geq 0 \quad \forall x,y \in \mathbb{R}$

$\therefore f$ is not convex

c) $\cos(x)$ is periodical $\Rightarrow$ not convex

3. $f(x,y) = 2xy^3 - 3x^2 - 6xy - 1$

$$\nabla f(x,y) = \begin{pmatrix} 2y^3 - 6x - 6y \\ 6xy^2 - 6x \end{pmatrix} \qquad \nabla^2 f(x,y) = \begin{pmatrix} -6 & 6y^2-6 \\ 6y^2-6 & 12xy \end{pmatrix}$$

a) $\nabla f(x,y) = 0 \Leftrightarrow \begin{cases} 2y^3 - 6x - 6y = 0 \\ 6xy^2 - 6x = 0 \end{cases} \Leftrightarrow \begin{cases} x = \frac{1}{3}y^3 - y \\ x(6y^2-6) = 0 \end{cases}$

$$\Leftrightarrow \left\{ \begin{array}{l} \\ (\frac{1}{3}y^3 - y)(6y^2 - 6) = 0 \end{array} \right\} \quad \left\{ \begin{array}{l} - \\ y(\frac{1}{3}y^2 - 1)(6y^2 - 6) = 0 \end{array} \right\}$$

$$y^* = \{0, \pm\sqrt{3}, \pm 1\}$$

$P_1 = (0,0)$ $P_2 = (0, -\sqrt{3})$ $P_3 = (0, \sqrt{3})$ $P_4 = (\frac{2}{3}, -1)$ $P_5 = (-\frac{2}{3}, 1)$

Inserting these points in $\nabla^2 f$:

P_1 : $\nabla^2 f(0,0) = \begin{pmatrix} -6 & -6 \\ -6 & 0 \end{pmatrix}$ indefinite matrix $\Rightarrow$ saddle point
$H_1 \leq 0$ $H_2 \geq 0$

...

P_5 : $\nabla^2 f(-\frac{2}{3}, 1) = \begin{pmatrix} -6 & 0 \\ 0 & -8 \end{pmatrix}$ negative definite $\rightarrow$ $H_1 < 0$; $H_2 < 0$
$\hookrightarrow$ local maximum

Homework 12

1.

- True. ReLU not differentiable at $x = 0$
- False. For $a < 0$ it is not true $\rightarrow$ $\sigma(-1 \times 1) = 0 \neq -1\,\sigma(1)$
- True $\sigma(\lambda x + (1-\lambda)y) \leq \lambda\sigma(x) + (1-\lambda)\sigma(y)$
- True

Notes

1. Introduction

- Random variables
- $N(0,1)$ $\rightarrow$ subtract mean, divide by σ
  $\hookrightarrow$ $E(z) = 0$, $\text{Var}(z) = 1$
- $E(x)$, μ_X , $\text{Var}(X)$, σ_X
- Central Limit Theorem : $\frac{\bar{X} - \mu_X}{\sigma/\sqrt{n}} \sim N(0,1)$
- $\text{Var}(X) = E(X^2) - E(X)^2$

2. Regression Analysis

- point estimate vs interval estimate
- confidence interval: $Pr\left(\bar{X} - z_{(1-\frac{\alpha}{2})}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{(1-\frac{\alpha}{2})}\frac{\sigma}{\sqrt{n}}\right)$
- if σ unknown, use s and t-distribution
- $s \to \frac{1}{n-1}$
- Hypothesis testing: H_0 vs H_1, α, p-value
  - $p \leq \alpha$, reject H_0
  - $p > \alpha$, insufficient evidence
- z-test: known σ
- student-t: unknown σ, $df = n-1$
- One-sided vs Two-sided
- Error Types: I → FP ; II → FN
- t-Test 2 samples: $t = \frac{\bar{d} - \Delta_0}{s/\sqrt{n}}$; ex: equal mean grades in 2 exams, same students
- p-value: probability of observing the data under H_0.
  - ↳ small is bad → no correlation between H_0 and data
- t-Test 2 independent samples: (Welch's Test)

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}} \qquad t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- Linear regression: $Y = \beta_0 + \beta_1 x$

  OLS: ordinary least squares → $\min \sum_i e_i^2 = \min \sum_i (y_i - \hat{y}_i)^2$

- $\hat{\beta}_1 = \frac{cov(x,y)}{var(x)}$; $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

  ↳ $cov(x,y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n-1}$; $var(x) = s^2 = \frac{\sum_i (x_i - \bar{x})^2}{n-1}$

- RSS: residuals → $RSS = \sum_i (\hat{y}_i - y_i)^2$
- ESS: explained → $ESS = \sum_i (\hat{y}_i - \bar{y})^2$
- TSS = ESS + RSS
- $R^2 = \frac{TSS - RSS}{TSS} = \frac{ESS}{TSS}$
  - 1 good
  - 0 no linear rel between x, y
- Testing coficients: $t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{\sqrt{\frac{RSS}{var(x)} \cdot \frac{1}{n-2}}}$

- Multiple Linear Regression: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots = X\hat{\beta}$
- OLS estimation: $y = X\beta + \varepsilon$

  $RSS = e^T e$ ← minimize

  ↳ $\beta = (X^T X)^{-1} X^T \cdot y$