T01 - Statistics

1. a) 1) Choosing the test: (paired samples)

- same group of individuals before and after the treatment (tax increase) ⇒ related samples
- Assumption: Quantity of interest is normally distributed.
  BUT True variance is unknown.
- small sample size
  ↳ paired t-test

Gaussian
- σ^2 known → z-test
- σ^2 unknown (small sample size) → t-test

- one-sample: $\mu_D = 0$
- two-sample: $\mu_A = \mu_B$
  - ↳ paired
  - ↳ independent

2) $H_0: \mu_d = 0$ ($:= \mu_0$) $\qquad H_1: \mu_d \neq 0$

3) Paired t-test:

$$\bar{d} = \frac{1}{10} \cdot \sum_{i=1}^{10} d_i = -8.1 \qquad s_2 = \frac{1}{n-1} \sum_{i=1}^{10} (x_i - \bar{x})^2$$

↳ Bessel's correction → unbiased sample variance

$$T_0 = \frac{\bar{d} - \mu_0}{\sigma_d / \sqrt{n}} = \frac{-8.1}{7.593/\sqrt{10}} = -3.3734$$

p-value: smallest α (significance) at which H_0 can be rejected.

4) $t^c = t^c_{1-\alpha/2;d} = t^c_{0.975;9} = 2.262$ ($\alpha = 0.05$, $d = n-1$)

5) Rejection criteria $|T_0| > t^c_{0.975;9} = 2.262$

$|T_0| = 3.3734 > 2.262 = t^c \Rightarrow H_0$ can be rejected!

6) Sufficient evidence that a tax increase has an effect on consumption!

2 a)

1) Type of test ; independent samples

- independent → Welch-test $\quad t_0 = \frac{\bar{x} - \bar{w} - N_0}{s_{\bar{x}-\bar{w}}} \sim t_{df}$

2) Hypothesis $\quad H_0 : \mu_f \leq \mu_m$ (status-quo) $\quad H_1 : \mu_f > \mu_m$ (claim)

$\Leftrightarrow H_0 = N_D := \mu_f - \mu_m \leq 0 \; : \; H_1 = N_D = \mu_f - \mu_m > 0$

3) Test statistics

$$T_0 = \frac{\bar{f} - \bar{m} - N_0}{s_{\bar{f}-\bar{m}}} = \frac{4 - 2.5 - 0}{0.769} \approx 1.949$$

- $\bar{f} = \frac{1}{10} \cdot \sum_{i=1}^{10} f_i = 4$
- $\bar{m} = 2.5$
- $s_{\bar{f}}^2 = \frac{1}{9} \cdot \sum_{i=1}^{10} (f_i - 4)^2 \approx 3.778$
- $s_{\bar{m}}^2 \approx 1.714$

$$s_{\bar{f}-\bar{m}} = \sqrt{\frac{s_{\bar{f}}^2}{10} + \frac{s_{\bar{m}}^2}{8}}$$

4) $H_0 : N_D := \mu_f - \mu_m \leq 0 \; : \; H_1 : N_D = \mu_f - \mu_m > 0$

$t^c = t^c_{1-\alpha;d}$ (one-sided) $= t^c_{0.95;16} = 1.746$

5) Rejection criteria $\quad |T_0| > t^c_{0.95;16} = 1.746$

$T_0 = 1.949 > 1.746 \Rightarrow H_0$ can be rejected!

6.) We reject H_0, which means we acknowledge that

H_1 : "On average, women wear their mask longer per day then men."

3. a) $H_0 : \mu_{NT} \leq \mu_{OT}$ $\quad H_1 : \mu_{NT} > \mu_{OT}$ ← we want to prove that NT improves results

$\Leftrightarrow H_0 : N_D := N_{NT} - N_{OT} \leq 0 \quad : \quad H_1 : N_D > 0$

b) $H_0 : N_D = N_0 = 0 \Rightarrow$ two-sided (there are two sides (higher and lower) the can violate the hypothesis)

$H_0 : N_D \leq 0 \Rightarrow$ One-side (only one side can violate (greater))

TO2 - Regression

1.

a) $\hat{\beta}_1 = \frac{cov(x,y)}{var(x)} = \frac{\frac{1}{8}((0.5-3.78)(5-63)+(0.6-3.78)(28-63)+\dots)}{\frac{1}{8}((0.5-3.78)^2+(0.6-3.78)^2+\dots)} = \frac{772.300}{133.776}$

$= 5.773$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 x = 41.178$

$\therefore \hat{y}(x) = 41.18 + 5.77x$

b) β_0 expected mean value for y when $x = 0$
β_1 expected change in y per unit of change in x

$\hat{\beta}_0$: A country with a capita gross national product of 0 has (approximately) 41.19% of literate people among the population

$\hat{\beta}_1$: with each increase of \$1.000, the percentage of literate people among the population increases by (approximately) 5.77.

c) Last-weeks:

1. Not needed, because only one test is used on regression analysis
2. $H_0: \beta_1 \leq 0$ vs $H_1: \beta_1 > 0$
3. Simply t-test for coefficient. We have

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

$$RSS \approx 4,411.4 \qquad \sum_{i=1}^{n}(x_i - \bar{x})^2 \approx 133.77$$

and thus $t_0 \approx 2.66$

4. $\alpha = 0.05$ (n - (number of coefficients))
5. For $df = n - 2 = 7$ degrees of freedom, we find a critical value of $t^c_{7;0.95} = 1.895$ in the t-table.
6. We reject H_0 because $t_0 > t^c$ and conclude that $\hat{\beta}_1$ is statistically significant.

d) Percentage of literate people (Y) might be over 100 when the capita gross national product (X) gets bigger than 13

We have $\hat{y} = 41.18 + 5.77x$

$X: 13.0 \rightarrow Y: 99$ New Zealand

$X: 14.0 \rightarrow 121.96$

T03 - Logistic Regression

2. a) $$L(y \mid x, \beta) = \prod_{i=1} p_i^{y_i} (1-p_i)^{1-y_i}$$

$$= p_3 \, p_4 \, (1-p_3)(1-p_2)$$

b) $$LL(\beta) = \sum_{i=1} y_i \log p_i + (1-y_i) \log (1-p_i)$$

$$LL(\beta) = \log p_3 + \log p_4 + \log (1-p_3) + \log (1-p_2)$$

$$\nabla LL(\beta) = \begin{pmatrix} \frac{\partial LL(\beta)}{\partial \beta_0} \\ \frac{\partial LL(\beta)}{\partial \beta_1} \end{pmatrix}$$

$$\frac{\partial LL}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial LL}{\partial p_i} \cdot \frac{\partial p_i}{\partial z_i} \cdot \frac{\partial z_i}{\partial \beta_j}$$

with $z_i = \beta_0 + \beta_1 x_i$ and $p_i = \sigma(z_i)$

$\sigma'(z_i) = \sigma(z_i)(1-\sigma(z_i))$

$$\frac{\partial LL}{\partial p_i} = \frac{y_i}{p_i} - \frac{1-y_i}{1-p_i} \qquad \frac{\partial p_i}{\partial z_i} = \sigma(z_i)(1-\sigma(z_i))$$

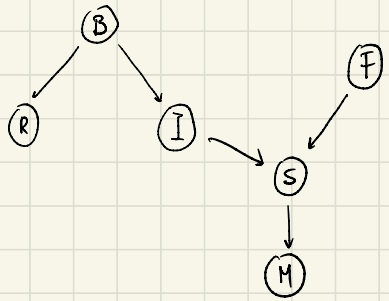
$$\rightarrow \frac{\partial LL}{\partial \beta_0} = \sum_{i=1} \left(\frac{y_i}{p_i} - \frac{1-y_i}{1-p_i} \right) \cdot \sigma(z_i)(1-\sigma(z_i)) \cdot 1$$

$$\frac{\partial LL}{\partial \beta_1} = \sum_{i=1} \left(\frac{y_i}{p_i} - \frac{1-y_i}{1-p_i} \right) \cdot \sigma(z_i)(1-\sigma(z_i)) \cdot x_i$$

c)

TO4 - Naive Bayes

1. a)



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

2. nominal: discr. without order — interval: cont. no zero
ordinal: discr. with order — ratio: cont. with an absolute 0

	cont / disc	range	scale of measurment
age	disc.	$(0, +\infty)$	ratio, ordinal
pref	disc.	{R, B, C}	nominal
money	cont.	$(0, +\infty)$	ratio
product	disc.	{UC, VC, NP}	nominal

TO5 - Decision Trees

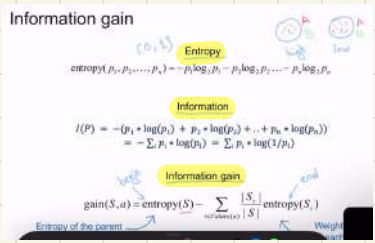
low since there is a 10%/90% division which is good and very pure.

1. a) entropy $((0.1, 0.9)) = -0.1 \log 0.1 - 0.9 \log 0.9 = 0.469$

b) entropy $((0.8, 0.2)) = -0.8 \log 0.8 - 0.2 \log 0.2 = 0.722$

c) entropy $((0.5, 0.5)) = -0.5 \log 0.5 - 0.5 \log 0.5 = 1$ → all classes have the same probability, the node is very impure → high entropy

d) entropy $((0.8, 0.1, 0.1)) = -0.8 \log 0.8 - 0.1 \log 0.1 - 0.1 \log 0.1 =$



e) info ([2,3]) = entropy $(\frac{2}{5}, \frac{3}{5}) = -\frac{2}{5}\log(\frac{2}{5}) - \frac{3}{5}\log(\frac{3}{5}) \approx 0.971$

f) info ([5,4]) = entropy $(\frac{5}{9}, \frac{4}{9}) \approx 0.991$

g) info ([2,3], [5,4]) = $\frac{5}{14}$ info ([2,3]) + $\frac{9}{14}$ info ([5,4]) ≈ 0.984

h) info ([2,3], [9,0]) = $\frac{5}{14}$ info ([2,3]) + $\frac{9}{14}$ info ([9,0]) ≈ 0.347 (0 → pure split)

2.

a) Only two possible splits : $x \leq 35$ or $x \leq 37$

35	35	37	40	40	40
F	F	T	F	T	T

35	35	37	40	40	40
F	F	T	F	T	T

1. info ([0,2], [3,1]) = 0.541 ← picked! (purest)

2. info ([1,2], [2,1]) ≈ 0.918

IG = info ([3,3]) − 0.541 ≥ info ([3,3]) − 0.918

↳ information of the original set

b)

1.		2.		3.	
36.8	36.8	37.2	38.3	38.3	39.7
T	F	F	T	F	F

info ([2,4]) = 0.918

1. info ([1,1], [1,3]) ≈ 0.874

2. info ([1,2], [1,2]) ≈ 0.918

3. info ([2,3], [0,1]) ≈ 0.809 ← picked!

∴ Since entropy of the original set is the same, for the highest IG (IG = entropy (S) − entropy (split)) we only need the lowest entropy split.

3. a) Past Trend

pos: p = 4 (up) n = 2 (down) P: $\frac{4}{6}$ N: $\frac{2}{6}$ Gini Index: $1 - P^2 - N^2 = 1 - \frac{4}{9} - \frac{1}{9} - \frac{4}{9}$

neg: p = 0 n = 0 P: 0 N: 1 Gini Index: 1 − 0 − 1 = 0

∴ Gini Index (past trend) = $\frac{6}{10} \cdot \frac{4}{9} + \frac{4}{10} \cdot 0 = \frac{4}{15}$

Gini Index (open interest) = $7/15$

Gini Index (trading volume) = $12/35$

b) past trend → lower Gini Index → purer nodes

T10 - PCA

1. a) $D = \{(-3,-1,-1), (0,-1,0), (-2,-1,2), (1,-1,3)\}$

1. Calculate the means $\quad \bar{d}_1 = \frac{-3+0+(-2)+1}{4} = \frac{-4}{4} = -1 \quad \bar{d}_2 = -1 \quad \bar{d}_3 = 1$

2. Subtract means

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

b) $\text{cov}(x_1, x_1) = \frac{1}{N-1} \sum_{i=1}^{N} x_{i,1}^2 = \frac{1}{3}(4+1+1+4) = 10/3$

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

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

$\text{cov}(x_2, x_2) = \frac{1}{3}(0) = 0 \qquad \text{cov}(x_2, x_3) = 0$

$\text{cov}(x_3, x_3) = \frac{1}{3}(4+1+1+4) = 10/3$

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

c) $|\Sigma_X - \lambda I| = 0 \Leftrightarrow \begin{bmatrix} 10/3-\lambda & 0 & 2 \\ 0 & -\lambda & 0 \\ 2 & 0 & 10/3-\lambda \end{bmatrix} = 0$

$\Leftrightarrow (10/3-\lambda)(-\lambda)(10/3-\lambda) + (0\cdot 0\cdot 2) + (0\cdot 0\cdot 2) - (2\cdot(-\lambda)\cdot 2) - (0\cdot 0\cdot(10/3-\lambda)) - ((10/3-\lambda)\cdot 0\cdot 0)$

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

d) Consider all λ:

$\lambda_1 = 0$: $(\Sigma_X - \lambda_1 I)\cdot \vec{v} = 0 \Leftrightarrow \begin{bmatrix} 10/3 & 0 & 2 \\ 0 & 0 & 0 \\ 2 & 0 & 10/3 \end{bmatrix} \cdot \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} = 0$

$\Leftrightarrow \begin{cases} 10/3\, v_1 + 2 v_3 = 0 \\ 2 v_1 + 10/3\, v_3 = 0 \end{cases} \Rightarrow v_1 = v_3 = 0$ $\quad v_2$ can take any value

normalize $|\vec{v}_1| = 1 \Rightarrow \vec{v}_1 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$

$\lambda_2 = 4/3$: $\vec{v}_2 = \begin{pmatrix} \sqrt{2}/2 \\ 0 \\ -\sqrt{2}/2 \end{pmatrix}$ $\qquad$ $\lambda_3 = 16/3$: $\vec{v}_3 = \begin{pmatrix} \sqrt{2}/2 \\ 0 \\ \sqrt{2}/2 \end{pmatrix}$

$\Phi = \begin{pmatrix} \sqrt{2}/2 & \sqrt{2}/2 & 0 \\ 0 & 0 & 1 \\ \sqrt{2}/2 & -\sqrt{2}/2 & 0 \end{pmatrix}$ $\quad \vec{v}_3 \quad \vec{v}_2$

For $\lambda_1 = 0$: $\frac{0}{0+\frac{4}{3}+\frac{16}{3}} = 0\%$

$\lambda_2 = \frac{4}{3}$: $\frac{4/3}{0+\frac{20}{3}} = 20\%$

$\lambda = \frac{16}{3}$: $\frac{16}{20} = 80\%$

e) $Z = X \cdot \Phi = \begin{bmatrix} -2 & 0 & -2 \\ 1 & 0 & -1 \\ -1 & 0 & 1 \\ 2 & 0 & 2 \end{bmatrix} \cdot \begin{pmatrix} \sqrt{2}/2 \\ 0 \\ \sqrt{2}/2 \end{pmatrix} = \begin{bmatrix} -2\sqrt{2} \\ 0 \\ 0 \\ 2\sqrt{2} \end{bmatrix}$

f) $Z = X \cdot \Phi = X \cdot \begin{pmatrix} \sqrt{2}/2 & \sqrt{2}/2 \\ 0 & 0 \\ \sqrt{2}/2 & -\sqrt{2}/2 \end{pmatrix} = \begin{bmatrix} -2\sqrt{2} & 0 \\ 0 & \sqrt{2} \\ 0 & -\sqrt{2} \\ 2\sqrt{2} & 0 \end{bmatrix}$

2. a) $D \approx Z \cdot \bar{\Phi}^T + \text{means}$

$\bar{d}_1$ $\bar{d}_2$ $\bar{d}_3$

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

b) $D \approx \begin{bmatrix} -2\sqrt{2} & 0 \\ 0 & \sqrt{2} \\ 0 & -\sqrt{2} \\ 2\sqrt{2} & 0 \end{bmatrix} \cdot \begin{bmatrix} \sqrt{2}/2 & 0 & \sqrt{2}/2 \\ \sqrt{2}/2 & 0 & -\sqrt{2}/2 \end{bmatrix} + \begin{bmatrix} -1 & -1 & 1 \\ -1 & -1 & 1 \\ -1 & -1 & 1 \\ -1 & -1 & 1 \end{bmatrix} = \begin{bmatrix} -3 & -1 & -1 \\ 0 & -1 & 0 \\ -2 & -1 & 2 \\ 1 & -1 & 3 \end{bmatrix}$

completely reconstructed since $\lambda_3 + \lambda_2$ explain 80% + 20% of the data alone

3. a)
1. Zero Mean Dataset
2. Covariance
3. Eigenvalues
4. Eigenvectors
5. Projection Matrix

T11 - Convex Optimization

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

$f(x)$ is convex iff $\forall x_1, x_2 \in \text{dom}(f)$

$\Rightarrow f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2) \quad \forall \lambda \in [0,1]$

if f is first-order differentiable, then ⇕

$\Rightarrow f(x_2) \geq f(x_1) + \nabla f(x_1)^T (x_2 - x_1) \quad \forall x_1, x_2 \in \text{dom}(f)$

f is second-order differentiable $\nabla^2 f(x) \geq 0 \quad \forall x \in \text{dom}(f)$

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

$$\nabla^2 f(x,y) = \begin{pmatrix} 9a\exp(3x) & b/2 \\ b/2 & 2 \end{pmatrix} \geq 0$$

$\lambda_1, \lambda_2 \quad \lambda_1 \geq 0 \quad \lambda_2 \geq 0 \iff$ $\text{Tr}(\nabla^2 f(x,y)) = \lambda_1 + \lambda_2 \geq 0$
$\text{Det}(\nabla^2 f(x,y)) = \lambda_1 \cdot \lambda_2 \geq 0$

Note: for every x, $\exp(3x) \geq 0$

$\text{Tr}(\nabla^2 f(x,y)) = 9a\exp(3x) + 2 \geq 0 \Rightarrow a \geq \frac{-2}{9\exp(3x)}$

$\text{Det}(\nabla^2 f(x,y)) = 18\exp(3x) - \frac{1}{4}b^2 \geq 0 \Rightarrow b \leq \pm\sqrt{4 \times 18\exp(3x)}$

2. • $h_1(x) = g_1(y) = g(Ax+b)$

for $h_1(x)$, $\forall x_1, x_2$ s.t. $Ax_1 + b$ and $Ax_2 + b \in \text{dom}(g_1)$

$$\begin{aligned} \Rightarrow h_1(\lambda x_1 + (1-\lambda)x_2) &= g_1(A(\lambda x_1 + (1-\lambda)x_2) + b) \\ &= g_1(\lambda(Ax_1 + b) + (1-\lambda)(Ax_2 + b)) \quad \text{) =} \\ &\leq \lambda g_1(Ax_1 + b) + (1-\lambda) g_1(Ax_2 + b) \quad \text{) convexity of } g_1 \end{aligned}$$

$= \lambda h_1(x_1) + (1-\lambda) h_1(x_2)$ ← def of h_1

- $h_2(\lambda x_1 + (1-\lambda)x_2) = c_1 g_1(\lambda x_1 + (1-\lambda)x_2) + c_2 g_2(\lambda x_1 + (1-\lambda)x_2)$ ← convexity of g_1, g_2
  $\leq c_1(\lambda g_1(x_1) + (1-\lambda) g_1(x_1)) + c_2(\lambda g_2(x_1) + (1-\lambda) g_2(x_1))$
  $= \lambda [c_1 g_1(x_1) + c_2 g_2(x_1)] + (1-\lambda)[c_1 g_1(x) + \dots$
  $= \lambda h_2(x_1) + (1-\lambda) h_2(x_2)$

- $h_3(\lambda x_1 + (1-\lambda)x_2) \overset{\text{def}}{=} \max\{ g_1(\lambda x_1 + (1-\lambda)x_1),\ g_2(\lambda x_1 + (1-\lambda)x_2)\}$
  $\leq \max\{ \lambda_1 g_1(x_1) + (1-\lambda) g_1(x_2),\ \lambda_2 g_2(x_1) + (1-\lambda) g_2(x_2)\}$
  $\leq \max\{ \lambda h_3(x_1) + (1-\lambda) h_3(x_2),\ \lambda h_3(x_1) + (1-\lambda) h_3(x_2)\}$

$g_1(x) \leq h_3(x)$
$g_2(x) \leq h_3(x)$

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

$\nabla f = \begin{pmatrix} \partial f/\partial x \\ \partial f/\partial y \end{pmatrix} = \begin{pmatrix} 4x - 2y - 3 \\ y - 2x - 1 \end{pmatrix}$

line-search: $\alpha^{(n)} = \underset{\alpha}{\text{argmin}}\, f(x^{(n)} - \alpha \cdot \nabla f(x^{(n)}))$

↳ minimizes the result of f in order to iteratively approach a local minimum

$\nabla f(0,0) = \begin{pmatrix} -3 \\ -1 \end{pmatrix}$ $\qquad f(0,0) = 5$

$\alpha^{(1)} = \underset{\alpha}{\text{argmin}}\, f(z^{(1)} - \alpha \cdot \nabla f(z^{(1)}))$
$= \underset{\alpha}{\text{argmin}}\ 12.5\alpha^2 - 10\alpha + 5$
$= 10/25 = 0.4$

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

$\alpha^{(2)} = \underset{\alpha}{\text{argmin}}\ f(z^{(2)} - \alpha \cdot \nabla f(z^{(2)}))$
$= \underset{\alpha}{\text{argmin}}\ 12.5\alpha^2 - 10\alpha + 3$
$= 10/25 = 0.4$

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

T12 - Neural Networks

1. 1. • $K = 1 \Rightarrow F(x) = W_1 x$ → linear with no bias TRUE

• $K = 2$ $\nabla l(F) = \frac{\partial l}{\partial F} \cdot \frac{\partial F}{\partial x} = -\frac{2}{n} \sum_{i=1}^{n} (y_i - F(x_i)) \cdot (-W_1 W_2)$

$= (W_1 W_2) \cdot \frac{2}{n} \sum_{i=1}^{n} (y_i - F(x_i))$

TRUE

• Increasing $K \Rightarrow F(x) = W_K W_{K-1} \dots W_1 x$

↳ all linear transformations can be mapped to $W_J = W_K W_{K-1} \dots W_1$

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. $K = 3$ W_3 scalar

$F(x) = W_3 W_2 W_1 x$ $\quad l_x = (y - F(x))^2$

$\frac{\partial l_x}{\partial W_3} = \frac{\partial l}{\partial F} \cdot \frac{\partial F}{\partial W_3} = -2 (y - F(x)) \cdot W_2 W_1 x$

2. a) $x \in \mathbb{R}^{2 \times 1}$ $\quad W^{[1]} \in \mathbb{R}^{2 \times 2}$ $\quad W^{[2]} \in \mathbb{R}^{2 \times 1}$

$b^{[2]} \in \mathbb{R}^{1 \times 2}$ $\quad b^{[2]} \in \mathbb{R}^{1 \times 1}$

Trainable parameters: $2 \times 2 + 2 + 2 \times 1 + 1 = 9$

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

$b^{[2]} \in \mathbb{R}^{1 \times 2} \longrightarrow \vec{1} \times b^{[2]} \in \mathbb{R}^{n \times 2}$ $\quad [\cdot\ \cdot] \rightarrow \begin{bmatrix} \cdot & \cdot \\ \vdots & \vdots \\ \cdot & \cdot \end{bmatrix}$

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

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

$$X^{[1]} = \sigma(z^{[1]}) = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.731 \\ 0.731 & 0.5 \\ 0.731 & 0.731 \end{pmatrix} \qquad \sigma = \frac{1}{1+e^{-x}}$$

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

$$X^{[2]} = \sigma(z^{[2]}) = \begin{pmatrix} 0.5 \\ 0.443 \\ 0.557 \\ 0.5 \end{pmatrix} = \hat{Y} \qquad Y = \begin{pmatrix} 0 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

$$\therefore L(y, \hat{y}) = \frac{1}{4} \sum_{i=1}^{n} l(Y_i, \hat{Y}_i) = 0.696$$

d) $\frac{\partial L}{\partial w^{[2]}} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z^{[2]}} \cdot \frac{\partial z^{[2]}}{\partial w^{[2]}}$

$\frac{\partial L}{\partial b^{[2]}} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z^{[2]}} \cdot \frac{\partial z^{[2]}}{\partial b^{[2]}}$ (last factor $= 1$)

$\frac{\partial L}{\partial w^{[1]}} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z^{[2]}} \cdot \frac{\partial z^{[2]}}{\partial x^{[1]}} \cdot \frac{\partial x^{[1]}}{\partial z^{[1]}} \cdot \frac{\partial z^{[1]}}{\partial w^{[1]}}$

$\frac{\partial L}{\partial b^{[1]}} = " \cdot " \cdot " \cdot " \cdot \frac{\partial z^{[1]}}{\partial b^{[1]}}$