Minor 1

Total marks: 15 pts

Total time: 45 mins

Instructions:

- **Write your name and roll number on answer script**
- With the exception of Question 1, all your answers must be clearly motivated! *A correct answer without a proper motivation will score zero points*!
- Vectors are denoted with a single underline $\underline{a}$, and matrices by double underline, $\underline{\underline{A}}$. Scalars appear without any underline. Please follow this rule through your answer book.

Some relevant formulas

- **The Gaussian distribution**: The probability density function of the p-dimensional Gaussian distribution with mean vector $\underline{\mu}$ and covariance matrix $\underline{\underline{\Sigma}}$ is

$$\mathcal{N}\left(\underline{x} \mid \underline{\mu}, \underline{\underline{\Sigma}}\right) = \frac{1}{(2\pi)^{p/2}\sqrt{\det \underline{\underline{\Sigma}}}} \exp\left(\frac{1}{2}\left(\underline{x} - \underline{\mu}\right)^T \underline{\underline{\Sigma}}^{-1}\left(\underline{x} - \underline{\mu}\right)\right)$$

- **Maximum likelihood**: The maximum likelihood estimate is given by

$$\hat{\underline{\theta}}_{\text{ML}} = \arg\max_{\underline{\theta}} \; p\left(y^{(1)}, \ldots, y^{(N)} \mid \underline{x}^{(1)}, \ldots, \underline{x}^{(N)}; \underline{\theta}\right)$$

where N is the number of training data points

- **Logistic regression**: The logistic regression combines linear regression with the logistic function to model the class probability

$$p(y = 1 \mid \underline{x}) = \frac{\exp\left(\underline{x}^T \underline{\theta}\right)}{1 + \exp\left(\underline{x}^T \underline{\theta}\right)}$$

For multi-class logistic regression, we use the *softmax* function,

$$p(y = m \mid \underline{x}) = \frac{\exp\left(\underline{x}^T \underline{\theta}^m\right)}{\sum_{j=1}^{M} \exp\left(\underline{x}^T \underline{\theta}^m\right)}$$

1. **[1 pt]** Answer `True` or `False`.
Each correct answer scores 0.25 point, each incorrect answer scores −0.25 point and each missing answer scores 0 point.

   (a) A classifier is called linear if the function that maps each input to a predicted class is linear in the parameters F
   (b) Normalizing the dataset is important for the performance of a decision tree F
   (c) Linear regression requires all input variables to be numerical in nature F
   (d) A company want to build a model for predicting the number of defective products manufactured during production. Since the number of defective products is an integer, this is best viewed as a classification problem F

2. **[1 pt]** Consider a case where we only measure two variables y and v, and we want to learn a linear regression model on a transformed input feature space. Can you identify the transformed input features for the following cases?

0.25 pt each

   (a) $y = \theta_0 + \theta_1 v + \theta_2 v^2 + \theta_3 v^3 + \theta_4 v^4 + \epsilon$ $x_1 = v, \quad x_2 = v^2, \quad x_3 = v^3, \quad x_4 = v^4$
   (b) $y = \theta_0 + \theta_1 v + \theta_2 \cos(v) + \theta_3 \sin(v) + \epsilon$ $x_1 = v, \quad x_2 = \cos(v), \quad x_3 = \sin(v)$
   (c) $y = \theta_0 + c\cos(v + \psi) + \epsilon$ (c and ψ are unknown constants) $\theta_0 + c\cos\varphi \underbrace{\cos v}_{x_1} + c\sin\varphi \underbrace{\sin v}_{x_2} + \epsilon$
   (d) $y = \theta_0 + \min\{\theta_1 v, \theta_2 v^2\} + \epsilon$ ← Not a linear regression model

3. **[2 pts]** Consider a scenario where you perform logistic regression on a dataset and you report training error and test error in terms of mis-classification percentage. Say the training error percentage is 20% and the test error percentage is 30% with logistic regression.

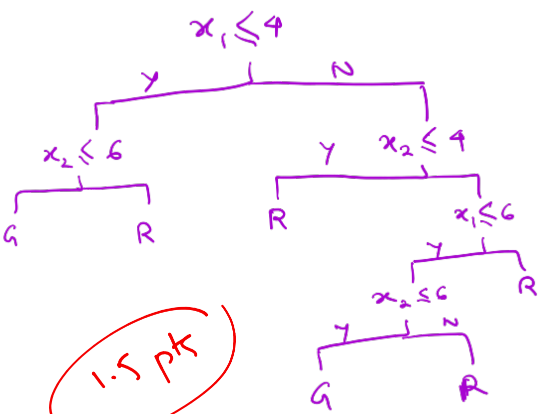
No marks without proper reason

Additionally, you try out k-NN (with $k = 1$), and you find the average of the training and test error percentage is 18%. Based on these results, which method should one use for classifying new unseen data and why? kNN (with k=1) has zero training error

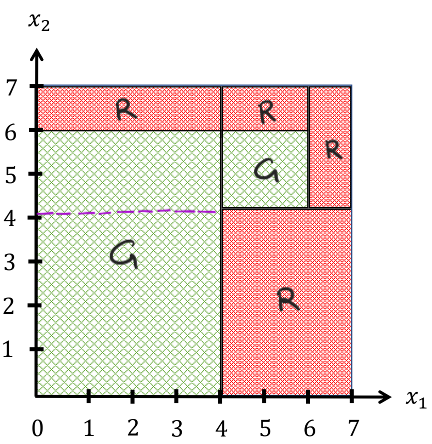
2 pts

[Use Logistic regression] So test-error for kNN = $\underline{2 \times 18\% = 36\%} \underline{>} 30\%$ (logistic regression)

4. **[1.5 pts]** Sketch (by hand) the classification tree corresponding for the following partition. How many leaves and internal nodes (including root node) does the resulting tree have?



1.5 pts



R - Red

G - Green

5. **[3 pts]** In class, linear regression using maximum likelihood (ML) approach was performed assuming Gaussian distribution with i.i.d. noise ϵ and the ML estimate came out to be $\underline{\hat{\theta}} = \left(\underline{\underline{X}}^T \underline{\underline{X}}\right)^{-1} \underline{\underline{X}}^T \underline{y}$.

   Consider the same linear regression model,

   $$y = \underline{x}^T \underline{\theta} + \epsilon.$$

   We are given N training data points $\left\{\underline{x}^{(i)}, y^{(i)}\right\}$, $i = 1, 2, \ldots, N$. However, in this case, the corresponding (unobserved) noise samples, $\epsilon^{(i)}$, $i = 1, 2, \ldots, N$, are assumed to follow a jointly Gaussian distribution with zero mean and covariance matrix equal to $\underline{\underline{\Sigma}}$. Derive a closed-form ML estimate of the parameters $\underline{\theta}$ for the correlated Gaussian noise case.

6. **[1.5 pts]** Derive logistic regression for binary classification as a special case of multi-class logistic regression.

7. **[4 pts]** Consider a scenario of estimating a constant, 5, using linear regression. Let's say we have a single measured data point $y^{(1)}$ ($N = 1$), which is generated from the true model $y = f_0(x) + \epsilon$, $f_0(x) = 5$. The noise ϵ has mean 0 and variance σ^2.

   We want to use linear regression with only one constant term θ, that is,

   $$y = \theta + \epsilon,$$

   and we learn θ using ridge regression with regularization parameter λ. The distribution $p(x)$ does not matter much here as there is no inputs x required.

   (a) **[1 pt]** Write out the closed-form solution for θ as a function of the training data $\mathcal{T} = \left\{y^{(1)}\right\}$ and the regularization parameter λ?

   (b) **[0.1 pt]** What is the expression for prediction $\hat{y}\left(x^*; \mathcal{T}\right)$

   (c) **[0.5 pt]** What is the average trained model $\bar{f}(x) = \mathbb{E}_{\mathcal{T}}\left[\hat{y}(x^*; \mathcal{T})\right]$? The expectation operator $\mathbb{E}_{\mathcal{T}}$ is an expectation over the training data.

   (d) **[0.5 pt]** What is the squared bias $\mathbb{E}_*\left[\left(\bar{f}(x^*) - f_0(x^*)\right)^2\right]$? The expectation operator $\mathbb{E}_*$ is an expectation over the test input $x^* \sim p(x)$. At what value of λ does the bias becomes minimum?

   (e) **[0.5 pt]** What is the variance $\mathbb{E}_*\left[\mathbb{E}_{\mathcal{T}}\left[\left(\hat{y}(x^*; \mathcal{T}) - \bar{f}(x^*)\right)^2\right]\right]$? At what value of λ does the variance becomes minimum?

   (f) **[0.15 pt]** What is the irreducible error $\mathbb{E}_*\left[\mathbb{E}_{\mathcal{T}}\left[\epsilon^2\right]\right]$?

   (g) **[0.25 pt]** What is the $\bar{E}_{\text{new}} = \mathbb{E}_{\mathcal{T}}\left[\mathbb{E}_*\left[\left(\hat{y}(x^*; \mathcal{T}) - y^*\right)^2\right]\right]$ for this problem?

   (h) **[1 pt]** For which value of the regularization parameter λ does the $\bar{E}_{\text{new}}$ become minimum?

5)

0.5

$$
\begin{cases}
\underline{y} = \underline{\underline{X}}\,\underline{\theta} + \underline{\epsilon} \\[2ex]
\underset{N\times 1}{\underline{\epsilon}} \sim \mathcal{N}\left(\underset{N\times 1}{\underline{0}}, \underset{N\times N}{\underline{\underline{\Sigma}}}\right)
\end{cases}
\qquad
\underline{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(N)} \end{bmatrix}
$$

$$
\underline{\underline{X}} = \begin{bmatrix} x^{(1)T} \\ x^{(2)T} \\ \vdots \\ x^{(N)T} \end{bmatrix} \text{ or } \begin{bmatrix} \underline{1}^T \\ x^{(1)T} \\ x^{(2)T} \\ \vdots \\ x^{(N)T} \end{bmatrix}
$$

<u>likelihood</u>

$$p(\underline{y} \mid \underline{\underline{X}}; \underline{\theta}) = \mathcal{N}(\underline{\underline{X}}\,\underline{\theta}, \underline{\underline{\Sigma}})$$

<u>log-likelihood</u>

$$L(\underline{\theta}) = \ln p(\underline{y} \mid \underline{\underline{X}}; \underline{\theta})$$

$$= \ln\left\{ \frac{1}{(2\pi)^{N/2} |\underline{\underline{\Sigma}}|^{1/2}} \exp\left(-\frac{1}{2}(\underline{y} - \underline{\underline{X}}\,\underline{\theta})^T \underline{\underline{\Sigma}}^{-1} (\underline{y} - \underline{\underline{X}}\,\underline{\theta})\right)\right\}$$

$$= -\frac{N}{2}\log(2\pi) - \frac{1}{2}\log|\underline{\underline{\Sigma}}| - \frac{1}{2}(\underline{y} - \underline{\underline{X}}\,\underline{\theta})^T \underline{\underline{\Sigma}}^{-1} (\underline{y} - \underline{\underline{X}}\,\underline{\theta})$$

1

We want to maximize the log-likelihood, so set its derivative to zero.

0.5

$$\frac{\partial L(\underline{\theta})}{\partial \underline{\theta}} = \underline{\underline{X}}^T \underline{\underline{\Sigma}}^{-1} (\underline{y} - \underline{\underline{X}}\,\underline{\theta}) = 0$$

$$\Rightarrow (\underline{\underline{X}}^T \underline{\underline{\Sigma}}^{-1} \underline{\underline{X}})\underline{\theta} = \underline{\underline{X}}^T \underline{\underline{\Sigma}}^{-1} \underline{y}$$

$$\Rightarrow \boxed{\hat{\underline{\theta}}_{ML} = (\underline{\underline{X}}^T \underline{\underline{\Sigma}}^{-1} \underline{\underline{X}})^{-1} \underline{\underline{X}}^T \underline{\underline{\Sigma}}^{-1} \underline{y}}$$

1

6) Binory logistic regression

$$p(y=1 \mid \underline{x}) = \frac{e^{\underline{x}^T \underline{\theta}}}{1 + e^{\underline{x}^T \underline{\theta}}}$$ [Given in formula]

$$p(y=-1 \mid \underline{x}) = 1 - p(y=1 \mid \underline{x})$$

$$= \frac{1}{1 + e^{\underline{x}^T \underline{\theta}}}$$

0.25

Multiclass logistic (or softmax) regression

$$p(y=m \mid \underline{x}) = \frac{e^{\underline{x}^T \underline{\theta}^m}}{\sum_{j=1}^{M} e^{\underline{x}^T \underline{\theta}^j}}$$

Consider two classes $m = \{1, 2\}$

$$p(y=1 \mid \underline{x}) = \frac{e^{\underline{x}^T \underline{\theta}^1}}{e^{\underline{x}^T \underline{\theta}^1} + e^{\underline{x}^T \underline{\theta}^2}}$$

0.5

$$= \frac{e^{\underline{x}^T (\underline{\theta}^1 - \underline{\theta}^2)}}{1 + e^{\underline{x}^T (\underline{\theta}^1 - \underline{\theta}^2)}} = \frac{e^{\underline{x}^T \underline{\theta}}}{1 + e^{\underline{x}^T \underline{\theta}}}$$

$$p(y=2 \mid \underline{x}) = \frac{e^{\underline{x}^T \underline{\theta}^2}}{e^{\underline{x}^T \underline{\theta}^1} + e^{\underline{x}^T \underline{\theta}^2}}$$

0.5

$$= \frac{1}{1 + e^{\underline{x}^T (\underline{\theta}^1 - \underline{\theta}^2)}}$$

$$= \frac{1}{1 + e^{\underline{x}^T \underline{\theta}}}$$

Treat $\underline{\theta} = \underline{\theta}^1 \cdot \underline{\theta}^2$

0.25

7) Closed-form solution of ridge-regression

$$\underline{\hat{\Theta}} = \left(\underline{\underline{X}}^T \underline{\underline{X}} + \lambda \underline{\underline{I}}\right)^{-1} \underline{\underline{X}}^T \underline{y}$$

a) For this problem, we only have $x = 1$

$$\underline{\hat{\Theta}} = (1+\lambda)^{-1} y^{(1)} = \frac{y^{(1)}}{1+\lambda}$$

1

b) $\hat{y}(\underline{x}; T) = \underline{\hat{\Theta}}$ 0.1

c) $\bar{f}(\underline{x}) = \mathbb{E}_T[\hat{y}(\hat{x}; T)]$

$$= \mathbb{E}_T[\hat{\Theta}(T)] = \mathbb{E}_T\left[\frac{y^{(1)}}{1+\lambda}\right] = \mathbb{E}_T\left[\frac{5+\epsilon}{1+\lambda}\right]$$

0.5

$$= \frac{5}{1+\lambda} + \frac{\cancelto{0}{\mathbb{E}_T[\epsilon]}}{1+\lambda}$$

$$= \frac{5}{1+\lambda}$$

d) $$\text{Bias}^2 = \mathbb{E}_x\left[\left(\bar{f}(x^*) - f_0(x^*)\right)^2\right]$$

$$= \mathbb{E}_x\left[\left(\frac{5}{1+\lambda} - 5\right)^2\right] = \mathbb{E}_x\left[\left(\frac{\cancel{5} - \cancel{5} - 5\lambda}{1+\lambda}\right)^2\right]$$

0.5

$$= \frac{25\lambda^2}{(1+\lambda)^2}$$

0.5 At $\lambda = 0 \rightarrow$ bias becomes minimum

(e) $\text{Variance} = \mathbb{E}_*\left[\mathbb{E}_\mathcal{T}\left[\left(\hat{y}(x^*;\mathcal{T}) - \bar{f}(x^*)\right)^2\right]\right]$

0.5

$$= \mathbb{E}_*\left[\mathbb{E}_\mathcal{T}\left[\left(\hat{\theta} - \frac{5}{1+\lambda}\right)^2\right]\right]$$

$$= \mathbb{E}_*\left[\mathbb{E}_\mathcal{T}\left[\left(\frac{y^{(1)}}{1+\lambda} - \frac{5}{1+\lambda}\right)^2\right]\right]$$

$$= \mathbb{E}_*\left[\mathbb{E}_\mathcal{T}\left[\left(\frac{\cancel{5}+\epsilon-\cancel{5}}{1+\lambda}\right)^2\right]\right] = \frac{1}{(1+\lambda)^2}\mathbb{E}_*\left[\mathbb{E}_\mathcal{T}[\epsilon^2]\right]$$

$$= \frac{1}{(1+\lambda)^2}\mathbb{E}_*[\sigma^2]$$

$$= \frac{\sigma^2}{(1+\lambda)^2}$$

0.5 Variance becomes 0 $\leftarrow$ at $\lambda \to \infty$

(f) $\mathbb{E}_*\left[\mathbb{E}_\mathcal{T}[\epsilon^2]\right] = \sigma^2$ 0.15

(g) $\bar{E}_{new} = \text{Bias}^2 + \text{Variance} + \text{Irreducible error}$

$$= \frac{25\lambda^2}{(1+\lambda)^2} + \frac{\sigma^2}{(1+\lambda)^2} + \sigma^2$$ 0.25

(h) $$\frac{\partial \bar{E}_{new}}{\partial \lambda} = \frac{1}{(1+\lambda)^3}\left[50\lambda(1+\lambda) - 50\lambda^2 - 2\sigma^2\right]$$

$$= \frac{1}{(1+\lambda)^3}\left[50\lambda - 2\sigma^2\right] = 0$$ 1

$$\Rightarrow \boxed{\lambda = \frac{\sigma^2}{25}}$$