Assignment 1

STAT 541, Winter 2025

Instructor: Andrew McCormack, mccorma2@ualberta.ca
Due date: January 24th.

To receive full marks you must show your work for all derivations.

Problem 1: Weighted Loss Functions

Suppose we have a classification problem with classes $\mathcal{Y} = \{1, \dots k\}$ and let the feature space be $\mathcal{X} = \mathbb{R}^p$. Consider the weighted loss function $L_w(f(x), y) = w_y I(f(x) \neq y)$, where $I(\cdot)$ is an indicator function and $w_1, \dots, w_k \in [0, \infty)$.

(a) Compute the risk $R_w(f, P)$ of a classifier $f(x)$ under L_w for some fixed distribution P over $\mathcal{X} \times \mathcal{Y}$.

(b) Find an expression, as a function of the w_i and $\Pr(y = j|x)$, for the oracle classifier f^* that attains the minimal risk with respect to the loss L.(**Hint**: Similar to what we did in class, we can find f^* in a pointwise fashion where for every $x \in \mathcal{X}$ we try to find the optimal value of $f^*(x) \in \{1, \dots, k\}$.)

(c) Suppose that x and y are *independent* under P, $k = 3$ and that the weights are equal $1 = w_1 = \cdots = w_3$. Defining $\pi_i = \Pr(y = i)$, determine the oracle risks under three different settings: $\boldsymbol{\pi} = (0.1, 0.5, 0.4)$, $\boldsymbol{\pi} = (0.3, 0.4, 0.3)$, and $\boldsymbol{\pi} = (0.1, 0.1, 0.8)$. Which of these three settings results in the easiest and hardest classification problems in the sense of having the lowest and highest oracle risks?

Problem 2: Predictions for Gaussian Data

We will assume that

$$(y, x) \sim \mathcal{N}_{p+1}\left(\begin{bmatrix} \mu_y \\ \boldsymbol{\mu}_x \end{bmatrix}, \begin{bmatrix} \Sigma_{yy} & \boldsymbol{\Sigma}_{yx} \\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{bmatrix} \right) \tag{1}$$

follows a multivariate normal distribution where $y \in \mathbb{R}$ is a random variable and $x \in \mathbb{R}^p$ is a random vector. **Important fact**: The conditional distribution of y given x is $y|x \sim \mathcal{N}_1\big(\mu_y + \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}(x - \boldsymbol{\mu}_x), \Sigma_{yy} - \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}\big)$.

(a) Find the oracle predictor and the oracle prediction error under squared error loss. What kind of class of functions does this oracle predictor lie in?

(b) If the conditional distribution $p(y|x)$ was kept the same as above but $p(x)$ was changed so that $x^\top = (x_1, \dots, x_p)$ has $x_i \overset{i.i.d}{\sim} \text{Poisson}(\lambda_i)$, $i = 1, \dots, p$ what would the oracle predictor and prediction error be? Note that you may assume that $\text{Var}(x) = \boldsymbol{\Sigma}_{xx}$.

(c) If the covariance matrix of (y, x) in (1) has the form $\Sigma_{yy} = \tau^2$, $\boldsymbol{\Sigma}_{xx} = \sigma^2 \boldsymbol{I}_p$, $\boldsymbol{\Sigma}_{xy} = \alpha \mathbf{1}_p$ describe how the oracle prediction error changes as each of τ^2, σ^2, α are changed. Provide justification for why the prediction error changes in this way.

Problem 3: Empirical Risk Minimization

Recall that an ERM predictor is some function $\hat{f}$ in a pre-specified set of functions where $\hat{f}$ minimizes the empirical risk:

$$\hat{f} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \ \frac{1}{n} \sum_{i=1}^{n} L(f(x^{(i)}), y^{(i)}).$$

(a) Find an ERM predictor when L is the squared error loss, $\mathcal{X} = \mathbb{R}^p$, and $\mathcal{F} = \{f(x) = \beta_0 + \sum_{i=1}^{p} \beta_i x_i\}$ is the set of all linear functions.

(b) Find the ERM predictor when $L(f(x^{(i)}), y^{(i)}) = |f(x^{(i)}) - y^{(i)}|$ and $\mathcal{F} = \{f : f(x) = \mu\}$ is the set of all constant functions (these functions do not depend on the input x!). For simplicity you should assume that the $y^{(i)}$ are distinct. You may also assume that n is even or n is odd (whichever case you find easier). (**Hints:** When n is even the ERM predictor is not unique. Note also that the absolute value function is not differentiable at zero but has well-defined directional derivatives everywhere. If you find a point μ_* where the derivative of the loss function equals zero I will accept this as an answer.)

Problem 4: Multivariate Data and SVDs

Let $X \in \mathbb{R}^{n \times p}$ be a matrix where $n \geq p$. Recall that the (thin) SVD of $X = UDV^\top$ where $U \in \mathbb{R}^{n \times p}$ has orthonormal columns, $D \in \mathbb{R}^{p \times p}$ is diagonal and $V \in \mathbb{R}^{p \times p}$ is orthogonal.

(a) What are the eigenvalues of the symmetric matrices $XX^\top$ and $X^\top X$? Prove that these matrices are positive semidefinite. A positive semidefinite matrix is a symmetric matrix $A \in \mathbb{R}^{p \times p}$ where $v^\top A v \geq 0$ for all $v \in \mathbb{R}^p$.

(b) If $z \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ what is the value of $E(\|z - \boldsymbol{\mu}\|^2)$? (Here $\|x\|^2 = x^\top x = \sum_{i=1}^{p} x_i^2$ is the standard norm on $\mathbb{R}^p$)

R Setup: For R coding assessments you should include the relevant output as well as your code. To do this, I recommend knitting an Rmarkdown pdf document. You can then merge this pdf with LaTeXed or scanned, handwritten solutions to the other problems. If you have not done this before you can get started by:

1. Downloading R from CRAN here: https://mirror.csclub.uwaterloo.ca/CRAN/
2. Downloading RStudio here: https://posit.co/download/rstudio-desktop/
3. Opening RStudio and creating a new RMarkdown file for the assignment.
4. Installing relevant packages by typing `install.packages("packagename")` into the console. You can alternatively use the tabs *Tools* → *Install Packages*.
5. Typing your code for the assignment into code chunk(s).
6. Clicking on the *knit* button and selecting *knit to pdf* to produce a pdf of the code and output.
7. You will need a LaTeX distribution installed in order for the document to knit correctly. An easy way to do this is by typing `tinytex::install_tinytex()` into the R console.

Problem 5: R — Risk Simulations

Let $x \sim \mathcal{N}_1(0,1)$, $g(x) = x^4 - 5x^2 + 4$ and $y|x \sim \mathcal{N}_1(g(x), 1)$.

(a) Draw $n = 20$ i.i.d. observations from the above distribution. Fit a linear model with a degree $d = 4$ polynomial in x to this data. That is, the linear model has the form $\hat{f}(x) = \hat{\beta}_0 + \sum_{k=1}^{4} \hat{\beta}_k x^k$. Plot the estimated prediction function $\hat{f}(x)$ against $g(x)$ over the interval $x \in [-4, 4]$.

(b) Draw 10,000 new, independent, observations of $(\tilde{x}^{(i)}, \tilde{y}^{(i)})$ from the same distribution. Estimate the risk $R(\hat{f}, P)$ by $\frac{1}{10000} \sum_{i=1}^{10000} (\hat{f}(\tilde{x}^{(i)}) - \tilde{y}^{(i)})^2$.

(c) To estimate $E(R(\hat{f}, P))$ repeat the above process 300 times obtaining $\hat{f}_1, \ldots, \hat{f}_{300}$ where we estimate $E(R(\hat{f}, P)) \approx \frac{1}{300} \sum_{i=1}^{300} R(\hat{f}_i, P)$ and each $R(\hat{f}_i, P)$ is estimated as in (b).

(d) Repeat (c) for the two other linear models with $d = 2$ and $d = 6$. Which of the three models yields the lowest expected prediction error?

(e) Repeat (d) for $n = 10$. Which linear model has the lowest expected prediction error in this case?

Some useful R functions: `lm, predict, rnorm, mean, plot, lines.`

Aside: Don't be stingy with the number of samples you draw in your simulations! Far too often people only draw 100 or 200 samples when drawing 10,000 takes only half a second longer and produces much more accurate results.