Problem 0.

(i) Give examples of typical loss functions. In each case, state whether it is more appropriate for a regression or a classification problem.

(ii) Whats is the expression of Bayes classifier under the $0/1$ loss? What is the optimal regression function under a square loss?

(iii) Explain what empirical risk minimisation is.

(iv) Provide a sketch of typical squared bias, variance, training error, and test error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches.

(v) Explain why each of the four curves has the shape displayed in part *(iv)*.

(vi) What is the definition of the excess risk? Show that it can be decomposed into two terms, the first one representing the estimation error, and the second one the approximation error. Sketch these three terms on a single plot, as a function of model complexity.

(vii) Discuss fundamental differences between the traditional view and the modern approach of data science. Which one is more focused on prediction, and which one on inference? You may illustrate your discussion using linear regression or logistic regression.

Problem 1. *Baye's Risk*
Consider the prediction of a students performance in a course (pass/fail) when given a number of important factors. First, let $Y = 1$ denote a pass and let $Y = 0$ stand for failure. The observation X corresponds to the number of hours of study per week. This, in itself, is not a predictor of a students performance, and we would need more information about the students quickness of mind, health, and social habits, and so on. Assume that the conditional probability $r(x) = \mathbf{P}(Y = 1|X = x) = x/(x + c)$, for some $c > 0$.

(i) Write down the expression of Baye's classifier.

(ii) Show that the corresponding Baye's risk is given by

$$\mathcal{R}^* = \mathbf{E}_X \left(\frac{\min(c, X)}{c + X} \right) .$$

(iii) Compute Baye's risk if $X = c$ with probability one. What is the value of $\mathcal{R}^*$ if X is uniformly distributed on the interval $[0, 4c]$? on the interval $[0, 2c]$? What is your interpretation?

Problem 2. *Optimality under Absolute Loss*
Let $(X, Y) \sim \mathbf{P}_{X,Y}$, such that $\mathbf{E}|Y| < \infty$. Find the function f^* minimising the risk under the absolute loss

$$f^* := \arg\min_f \mathbf{E}\left\{|Y - f(X)|\right\},$$

where the minimum is over the space of all measurable functions.

Problem 3. Bias-Variance Decomposition
In this Problem we address the bias-variance tradeoff for the square loss function. The input variable is denoted by $x \in \mathcal{X}$, and the output variable by $y \in \mathcal{Y}$. A learning rule is a function $f : \mathcal{X} \to \mathcal{Y}$ which predicts the output y associated with the input x. We assume that $(x,\, y)$ is a realisation of a generic $(X,\, Y) \sim \mathbf{P}_{X,Y}$. Expectation under $\mathbf{P}_{X,Y}$ is denoted $\mathbf{E}_{X,Y}\{\cdots\}$. The estimation of y is based on a learning sample $\mathcal{L}_n := \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, and denoted $\hat{f}_n(x) = f(x,\, \mathcal{L}_n)$. For instance, $\hat{f}_n$ is the function minimising the empirical risk.

(i) Let f be a fixed learning rule. The risk of f for the square loss function is defined as

$$\mathcal{R}(f) := \mathbf{E}_{X,Y}\left\{\ell(Y, f(X))\right\} = \mathbf{E}_{X,Y}\left\{(Y - f(X))^2\right\}.$$

Show that we have the decomposition

$$\mathcal{R}(f) = \mathbf{E}_X\{\mathrm{var}(Y \,|\, X)\} + \mathbf{E}_X\left\{(f(X) - \mathbf{E}(Y \,|\, X))^2\right\}.$$

Explain in words what the first and second term on the right hand side represent.

(ii) We adopt a frequentist approach: uncertainty around the estimation of f is due to the randomness present in the data. The risk associated with $\hat{f}_n$ is

$$\mathcal{R}(\hat{f}_n) := \mathbf{E}_{X,Y}\left\{(Y - \hat{f}_n(X))^2 \,|\, \mathcal{L}_n\right\},$$

and the expected risk, taken over the distribution of $\mathcal{L}_n$, is $\mathbf{E}_{\mathcal{L}_n}\mathcal{R}(\hat{f}_n)$. Making use of the decomposition in *(i)*, show that for the square loss function, we have

$$\mathbf{E}_{\mathcal{L}_n}\mathcal{R}(\hat{f}_n) = \mathbf{E}_X\{\mathrm{var}(Y \,|\, X)\} + \mathbf{E}_X\left\{[\mathbf{E}_{\mathcal{L}_n}(\hat{f}_n(X) \,|\, X) - \mathbf{E}(Y \,|\, X)]^2\right\} + \mathbf{E}_{X,\mathcal{L}_n}\left\{[\hat{f}_n(X) - \mathbf{E}_{\mathcal{L}_n}(\hat{f}_n(X) \,|\, X)]^2\right\}.$$

Explain in words what these three terms represent.

Problem 4. *Baye's Classifier*
Consider a toy binary classification problem with output variable $Y \in \{0, 1\}$ and discrete conditional distributions of X (input variable) indicated in the following table:

x	1	2	3	4	5	6	7	8	9	10
$\mathbf{P}(X = x \,\|\, Y = 1)$	0.04	0.07	0.06	0.03	0.24	0	0.02	0.09	0.25	0.2
$\mathbf{P}(X = x \,\|\, Y = 0)$	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Bayes classifier										

Suppose that $\mathbf{P}(Y = 1) = 2/3$.

(i) What is Bayes classifier here and what is its error rate (for the 0/1 loss)? Fill in the table above with your predictions for each category.

(ii) We cannot observe x directly but only

$$x^* = \begin{cases} 2 & \text{if } x = 1 \text{ or } 2 \\ 4 & \text{if } x = 3 \text{ or } 4 \\ 6 & \text{if } x = 5 \text{ or } 6 \\ 8 & \text{if } x = 7 \text{ or } 8 \\ 10 & \text{if } x = 9 \text{ or } 10 \end{cases}$$

What is the optimal classifier and what is its error rate, again assuming that $\mathbf{P}(Y = 1) = 2/3$ and using 0/1 loss?