Problem 1 (On Errors and Models) **(10 points)**

(a) Consider Figure 1. Which of the following three options correctly describes what is happening in the figure? *You do not need to explain your answer.* (1 point)

i) Starting at *flexibility*= 1, with increasing flexibility, the increase in variance is smaller than the decrease in bias, resulting in the downward trend in the curve. As we increase flexibility over *flexibility*= 8, the variance starts to increase more rapidly than the bias decreases, hence causing a upward trend.

ii) Starting at *flexibility*= 1, with increasing flexibility, the increase in bias is smaller than the decrease in variance, resulting in the downward trend in the curve. As we increase flexibility over *flexibility*= 8, variance starts to increase more rapidly than the bias decreases, hence causing a upward trend.

iii) Starting at *flexibility*= 1, with increasing flexibility, the increase in bias is smaller than the decrease in variance, resulting in the downward trend in the curve. As we increase flexibility over *flexibility*= 8, bias starts to increase more rapidly than the variance decreases, hence causing a upward trend.

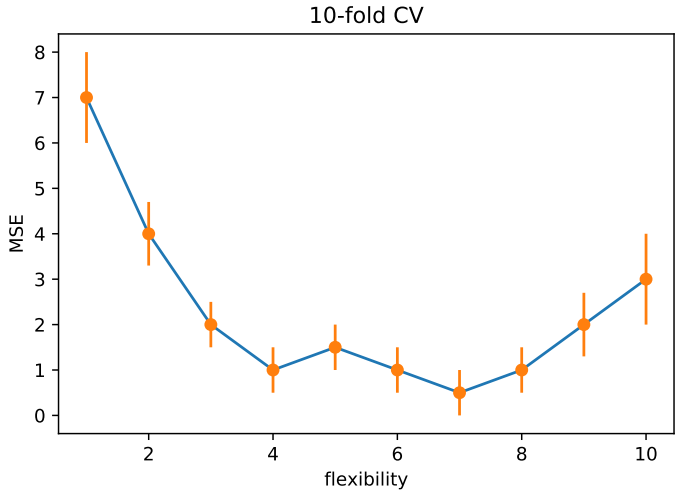

Figure 1: Test MSE for 10-fold Cross Validation (CV) for an unknown data set. Orange bars indicate the standard error.

(b) According to Figure 1, which model (flexibility), would you select? Explain why you chose this model. (2 points)

(c) State, for each of the three settings below, what will happen both in terms of bias and variance when we make the proposed change to the learning procedure. Indicate with a (+) if the given quantity increases, (–) if the given quantity decreases, and (=) if there is no change. Provide also an explanation, why the changes introduced in the specific model/method do (not) change the flexibility. (3 points)

Action	Bias	Variance	Flexibility	Explanation
1) Fitting data generated by $Y = \beta X + \mathcal{N}(0, 1)$ instead of $Y = \beta X + \mathcal{N}(0, 10)$.				
2) Changing from $K = 2$ to $K = 10$ in the K-nearest neighbor classifier.				
3) Setting budget C in the Support Vector Classifier to a higher value.				

(d) In linear regression, why do we *in general* assume that the error (noise) term averages to zero? (1 point)

(e) Describe in your own words the difference between a parametric and a non-parametric method. (1 point)

(f) Would you then consider the following methods parametric or non-parametric? Explain why each of the models fall under the definition of parametric/non-parametric. (2 points)

1) Logistic Regression
2) Decision Trees
3) Linear Discriminant Analysis
4) k-Nearest Neighbors (k-NN)

Solution.

(a) i)

(b) Model with *flexibility* = 4
One-standard-error rule: Choose the simplest model within one standard error of the best model (due to preference for more simpler models e.g. easier to interpret)
Model with *flexibility* = 7 has lowest MSE of all models. Out of all other models, model with *flexibility* = 4 is the model (i) whose MSE falls within one standard deviation of the MSE of Model *flexibility* = 7 and (ii) has least complexity.

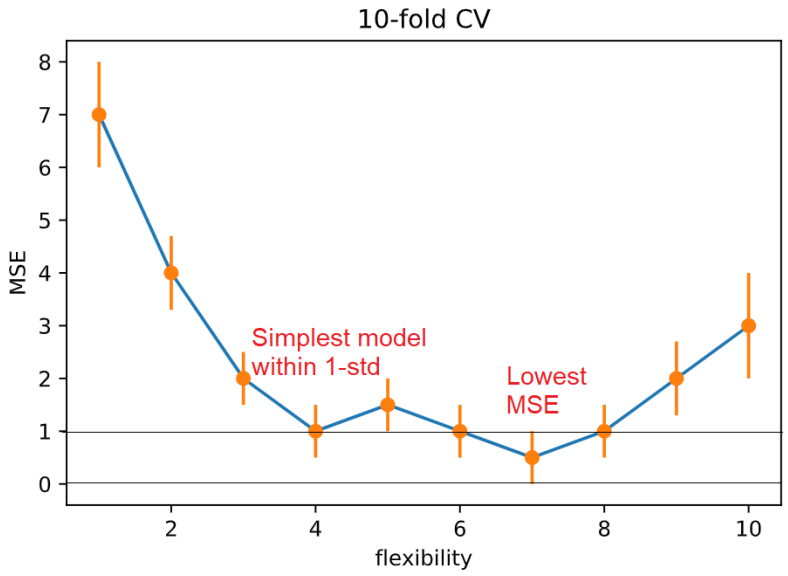

Figure 2: Illustration of the One-standard-error rule for the given scenario.

(c) Overview in the table, explanations below.

	Action	Bias	Variance	Flexibility
1)	Fitting on data generated by $Y = \beta X + \mathcal{N}(0,1)$ instead of $Y = \beta X + \mathcal{N}(0,10)$.	=	=	=
2)	Changing from $K = 2$ to $K = 10$ in the K-nearest neighbor classifier.	+	−	−
3)	Setting budget C in the Support Vector Classifier to a higher value.	+	−	−

Explanations:

1) We decrease irreducible error, which according to Expected MSE decomposition has no influence on Variance and Bias.

2) We increase the number of neighbours in majority voting, which fits our classifier less to the exact data point, but more to its surroundings.

3) With large C the margin is wide and there are many support vectors.

(d) Noise is assumed to be a random signal centered around the true value and not to introduce any additional bias.

(e) Parametric methods assume functional form with a fixed number of parameters. Non-parametric methods do not make explicit assumptions about the functional form of f. The numbers of parameters increases with amount of data.

(f) 1) Logistic Regression: parametric. Assumes functional form $p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$ with a fixed number of parameters

2) Decision Trees: non-parametric, does no make assumptions on the data, splitting in different numbers of regions with depth as hyper-parameter

3) Linear Discriminant Analysis: parametric, assumes distribution of the data to be Gaussian

4) KNN: non-parametric. Majority voting amongst neighbours for each point.

Problem 2 (Regression) **(15 points)**

You have been given data in Table 1 from a small experiment.

X_1	X_2	Y
-2	-3	-1
-2	-1	1
1	2	2
3	2	3

Table 1: Observations for predictor variable X_1 and X_2 and target variable Y.

(a) You want to perform univariate linear regression of predictor X_1 on response Y. Recall that simple linear regression takes the form $Y = \beta_1 \mathbf{X_1} + \beta_0$, but that it is often convenient to formulate it as $\mathbf{X}\beta = \mathbf{Y}$ with $\beta = \begin{bmatrix}\beta_1 \\ \beta_0\end{bmatrix}$ and $\mathbf{X} = [\mathbf{X_1}; \mathbb{1}]$. Using the following conversion,

$$(\mathbf{X^T}\ \mathbf{X})^{-1} = \begin{bmatrix}0.06 & 0 \\ 0 & 0.25\end{bmatrix},$$

find the least square estimators $\hat{\beta}_1$ and $\hat{\beta}_0$. Explain the reasoning behind each step. (3 points)

Assume now that we collect a thousand data points to predict the response variable Y using one million predictors.

(b) Is least squares linear regression appropriate in this setting? Explain your answer. (1 point)

(c) Assume we apply forward stepwise selection and the results indicate that 9 of these predictors lead to a good predictive model on a given training data set. Can we ensure that these 9 predictors are the optimal set, i.e., the set of 9 predictors with minimum MSE? Explain why (not)? (1 point)

We now want to perform Principal Component Analysis (PCA) for dimensionality reduction using the data in Table 1, i.e. we want to reduce the two dimensional data $X = [X_1, X_2]$ to one dimension.

(d) Compute the covariance matrix. Indicate all the steps that you follow to compute it. (The final result alone does not give any points.) *Hint:* The covariance matrix takes the form (2 points)

$$\begin{pmatrix}Cov(X_1, X_1) & Cov(X_2, X_1) \\ Cov(X_1, X_2) & Cov(X_2, X_2)\end{pmatrix}$$

(e) Now compute the first principal component using the covariance matrix computed before. Explain what you do in each step. *Hint:* (2 points)

$$det \begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - cb$$

If you did not solve part (a), you may use this covariance matrix:

$$\begin{pmatrix} 4.5 & 4 \\ 4 & 4.5 \end{pmatrix}$$

(f) Draw the first and second principal component in a plot. It is sufficient to indicate (i) the angle between the two principal components, (ii) the angle between each principal component and the x-axis and (iii) the module (length) of each principal component. (1 point)

(g) What important assumption does PCA make on the statistical relationship (i.e., dependence) between the predictor variables? (1 point)

(h) Do *in general* the LS estimator for the first predictor (e.g. $\hat{\beta}_1$ computed in a)) and the first principal component of PCA (e.g. computed in e)) point to the same direction? Explain why (not)? (1 point)

Suppose we have two data sets with the same $n = 500$ observations and the same $p = 20$ predictors, but with two different response variables Y_1 and Y_2. In data set 1 response variable Y_1 is a function of all the predictors, whereas in data set 2 the response variable Y_2 only depends on two of the predictors. We perform PCA for dimensionality reduction and extract the first five principle components for both the data sets.

(i) Will the principle components be the same for both data sets? Explain why (not)? (1 point)

(j) Next we perform Principal Component Regression (PCR) using the first five principle components that we extracted. For which data set would you expect to have a better accuracy? Explain why. (1 point)

(k) Assume you now also perform Partial Least Squares (PLS) regression with 5 principle components for both datasets. Which method, PCR or PLS do you expect to have the smaller training Residual Sum of Squares (RSS) for Y_1 and for Y_2? Explain why. (1 point)

Solution.

(a) (i) Give formula (1 P)

$$X^T X\beta = X^T Y$$
$$\beta = (X^T X)^{-1} X^T Y$$

(ii) Valid reason for using the formula, e.g. stating it is because we minimize RSS (1 P)

(iii) Multiply the given matrices to get the solution $\beta = (0.66, 1.25)$. Numerical errors overlooked in most cases. (1 P)

(b) No. When $p > n$ the $X^T X$ is no longer invertible, meaning that the problem is underdetermined, there are infinitely many possible solutions. Moreover, it will yield a set of coefficient estimates that result in a perfect fit to the data, this will almost certainly lead to overfitting of the data.

(c) No. Forward-stepwise selection begins with a model containing no predictors, and then adds predictors to the model, one-at-a-time, until all of the predictors are in the model. It is a greedy approach, in particular, at each step the variable that gives the greatest additional improvement to the fit is added to the model. It is not guaranteed to find the best possible model out of all models containing subsets of the p predictors. For instance, suppose that in a given data set with p = 3 predictors, the best possible one-variable model contains X_1, and the best possible two-variable model instead contains $X2$ and $X3$. Then forward stepwise selection will fail to select the best possible two-variable model, because it never *deselects* a predictor once it has entered the model, in this case X_1 (**see pg. 208 Sec. 6.1.2 ISLR**).

(d)

$$\begin{aligned}
\text{Cov}(X_1, X_2) &= E((x_1 - E(X_1)(X_2 - E(X_2)) \\
E(X_1) &= E(X_2) = 0 \\
\text{Cov}(X_1, X_1) &= \frac{1}{4}((-2)^2 + (-2)^2 + (1)^2 + (3)^2) \\
&= \frac{1}{4}(4 + 4 + 1 + 9) \\
&= \frac{18}{4} \\
&= 4.5
\end{aligned}$$

$$\begin{aligned}
\text{Cov}(X_1, X_2) &= \frac{1}{4}((-2)(-3) + (-2)(-1) + (1)(2) + (3)(2)) \\
&= \frac{1}{4}(6 + 2 + 2 + 6) \\
&= 4 \\
&= 4.5
\end{aligned}$$

$$\begin{pmatrix} Cov(X_1, X_1) & Cov(X_2, X_1) \\ Cov(X_1, X_2) & Cov(X_2, X_2) \end{pmatrix} = \begin{pmatrix} 4.5 & 4 \\ 4 & 4.5 \end{pmatrix}$$

(e) Step 1: Eigenvalues

$$\begin{pmatrix} 4.5 - \lambda & 4 \\ 4 & 4.5 - \lambda \end{pmatrix} = (4.5 - \lambda)^2 - 16 = 0$$

$$\lambda_1 = 0.5$$
$$\lambda_2 = 8.5$$

Step 2: First Eigenvectors - The eigenvectors represent the directions or components for the reduced subspace of B, whereas the eigenvalues represent the magnitudes for the directions. The first principal is the direction with highest eigenvalue, here λ_2.

$$\begin{pmatrix} 4.5 & 4 \\ 4 & 4.5 \end{pmatrix} \begin{pmatrix} v_1^{(2)} \\ v_2^{(2)} \end{pmatrix} = \begin{pmatrix} 8.5 \\ 8.5 \end{pmatrix}$$

Eigenvector: $v^{(2)} = (1, 1)$

(f)

	length	angle
First principal component:	8.5 or 1.0	45 °
Second principal component:	0.5 or 1.0	$v_1 = (-1, 1)$ then 315 °(45 °) or $v_1 = (1, -1)$ then 135 °
Between both principle components:		90 °

(g) linear correlation

(h) No. When maximizing the variance we minimize the sum of squared distances between the projected datapoint and the original datapoint computed on X_1 and X_2, while when fitting a linear function we minimize the squared distance between the predicted datapoint and the original value computed on X_1 and Y .

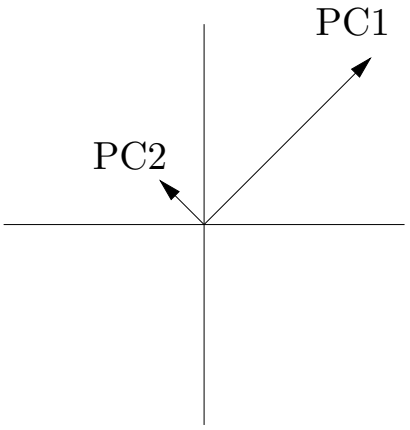

Figure 3: First and second principal component with $v_1 = (-1, 1), v_2 = (1, 1)$ and $\lambda_1 = 0.5, \lambda_2 = 8.5$.

(i) Yes, because we perform PCA only on X and not on Y. The datasets only differ on Y.

(j) It depends. We expect PCR to work well on predicting Y_1, i.e. on the response variable that depends on all predictors. The first five components are assumed to capture a great part of variance over all predictors and thus should be informative about Y_1. Regarding Y_2 it will depend on the variance of two predictors that Y_2 and whether they are accurately represented in the 5 first components. If the 5 first PCs do not contain information about the two predictors that Y_2 depends on, then the fit of Y_2 will be poor. On the contrary, if all information about this two predictors is in the 5 first PCs, then the fitting of Y_2 should be as accurate (or better) as for Y_1.

(k) PLS would work best for both datasets, as it accounts for the response variable (i.e., Y_1 or Y_2) when computing the principlal components.

Problem 3 (Classification) **(10 points)**

Consider now a binary classification problem, i.e. the label can assume either $(Y = 1)$ or $(Y = 0)$.

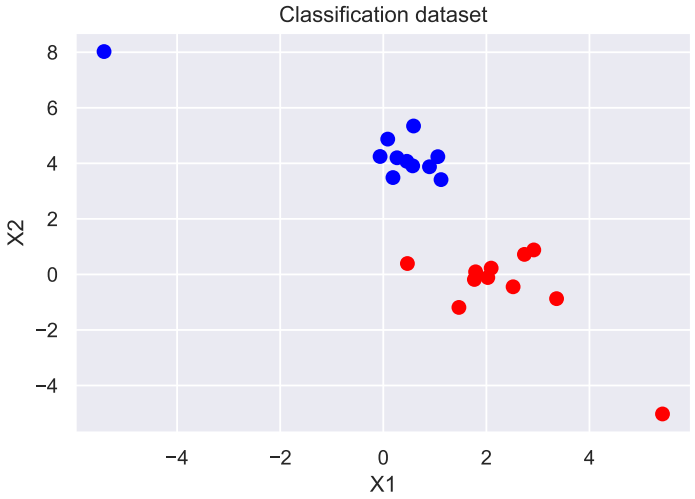

Figure 4: Classification data set with two classes.

Looking at Figure 4 you are wondering, if you should first pre-process the data to get rid of the outliers before training a classification model. To save time you first take a look at the different methods:

(a) Are the following methods sensitive to outliers, where outliers are seen as a set of predictors that are out of the predictor distribution (in Figure 4, the two observations in the top-left and bottom-right corners)? Explain why (not). (2 points)

1) Logistic Regression
2) Decision Trees
3) Support Vector Machines (SVM)
4) k-Nearest Neighbours (k-NN)

Let's take a closer look at Support Vector Machines (SVM). Recall that an SVM are defined as follow:

$$\underset{\beta_0,\ldots,\beta_p,\ \xi_1,\ldots,\xi_N}{\text{maximize}} \quad M \tag{3.1}$$

$$\text{subject to} \quad \|\beta\| = 1 \tag{3.2}$$

$$\xi_i \geq 0 \tag{3.3}$$

$$y_i f(x_i) \geq M(1-\xi_i) \quad \text{for } i = 1,\ldots,N \tag{3.4}$$

$$\sum_{i=1}^{N} \xi_i \leq C \tag{3.5}$$

(b) Describe the constraints (3.3) and (3.5) in your own words. Which different values can ξ_i take and what do the different values express? (2 points)

(c) Describe the purpose of constraints (3.2) and (3.4) in your own words. How are these two constraints related to each other? (3 points)

Now assume that there there would not be just two, but $K = 10$ classes. Let $f_k(x)$ denote the density function of X, i.e $\Pr(X = x \mid Y = k)$, for the observation that comes from the kth class. Recall, according to Bayes' Theorem: $\Pr(Y = k \mid X = x) \propto \pi_k \cdot f_k(x)$. You suspect your data to follow an Poisson distribution with distinct λ_k for each of the k classes, where

$$f(x;\lambda_k) = \frac{(\lambda_k)^x e^{-\lambda_k}}{x!}$$

(d) Derive the discriminant function, in a similar way that we did for a Gaussian likelihood for Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA). (2 points)

(e) Is the above derived discriminant function linear in terms of x? Explain why (not). (1 point)

Solution.

(a)
1) Logistic regression: No. Decision boundary takes into consideration only the points that are closer to it.
2) Decision Trees: No. Splitting is determined based on the sample proportions in each split region and not absolute values.
3) Support Vector Machine: No. The decision boundary in SVM only depends upon the support vectors. It does not 'care' about samples on the correct side of the margin at all.
4) K-Nearest Neighbours: No. It is based on majority voting, so the outcome will not be dominated by the outlier.

(b)
- Constraint (3.3): For each point i: Non-negativity of the distance ξ_i from the point to x. If $\xi_i > 1$ then the i-th observation is on the wrong side of the hyperplane, if $\xi_i > 0$ the ith observation violates the margin.
- Constraint (3.5): C specifies the total amount of slack allowed (budget), i.e. amount that the margin can be violated by the N observations. The sum over all points $i = 1, \ldots, N$ of all individual distances between x and y of all points i within the margin or on the wrong side of the margin needs to be smaller or equal to the total amount of slack allowed.

(c)
- Constraint (3.2): The norm of the parameters of the hyperplane that is used to classify points is said to be exactly 1. Ensures that the perpendicular distance from the i-th observation to the hyperplane is given by $y_i(\beta_0 + \beta_1 x_{i1} + ...\beta p x_{ip})$.
- Constraint (3.4): The LHS of this constraint allows us to measure the distance of a point from the hyperplane. The RHS of this constraints allows us to make sure that this distance does not always have to be greater than M, but in fact, can be also be smaller (to accommodate the points that we might misclassify).
  Relation between constraint (3.2) and (3.4): β are the parameters of $f(x_i)$. Classification occurs according to the sign of $f(x_i)$ and the magnitude of $f(x_i)$ indicates the confidence about the class assignment, i.e. if it is far from zero, then this means that x_i lies far away from the hyperplane. Constraint (3.4) requires each observation to be on the correct side of the hyperplane and outside of the margin, with a certain budget of slack indicated by ξ_i according to the budget C. Under constraint (3.2), the quantity $y_i f(x_i)$ gives the perpendicular distance from the i-th observation to the hyperplane. Therefore, the constraints (3.2) and (3.4) together ensure that each observation is on the correct side of the hyperplane and at least a distance M from the hyperplane. **(See pg. 342 Sec. 9.1.4 ISLR).**

(d) We know that $P(Y = k \mid X = x) \propto \pi_k \cdot f_k(x)$. We can plug-in the calculated π_k and the given $f_k(x)$ in to the equation of the Bayes Theorem. Then, we can approximate the Bayes optimal decision by taking $\arg\max_k$ of the product $\pi_k \cdot f_k(x)$.

$$
\begin{aligned}
\underset{k}{\text{argmax}} \;\; p_k(x) &= \underset{k}{\text{argmax}} \;\; \pi_k \, \frac{(\lambda_k)^x e^{-\lambda_k}}{x!} \\
&= \underset{k}{\text{argmax}} \;\; \log(\pi_k) + \log\left((\lambda_k)^x\right) + \log\left(e^{-\lambda_k}\right) - \log\left(x!\right) \\
&= \underset{k}{\text{argmax}} \;\; \log(\pi_k) + \log\left((\lambda_k)^x\right) - \lambda_k - \log\left(x!\right) \\
&= \underset{k}{\text{argmax}} \;\; \log(\pi_k) + x \log\left(\lambda_k\right) - \lambda_k - \log\left(x!\right) \\
&= \underset{k}{\text{argmax}} \;\; \log(\pi_k) + x \log\left(\lambda_k\right) - \lambda_k
\end{aligned}
$$

(e) This classifier is linear in terms of x, as we drop the term $\log(x!)$ in the last line of (d) since it is independent of k and we optimize for k (argmax_k).

Problem 4 (Beyond Linear Regression) **(5 points)**

Now assume we are interested in how a predictor X_1 relates to the response variable Y. Prior analysis shows that this relationship is non-linear. One modeling option would be to use regression splines.

(a) How many degrees of freedom does a regression spline have if we use polynomials of degree $d = 3$, have $K = 10$ knots, and require the spline to be continuous at the knots up to the first derivative. Explain your answer. (1 point)

(b) Up to which derivative do we have to enforce continuity at the knots if we want 25 degrees of freedom, polynomials of degree $d = 4$, and have $K = 10$ knots? Explain your answer. (1 point)

Let's now consider Smoothing Splines.

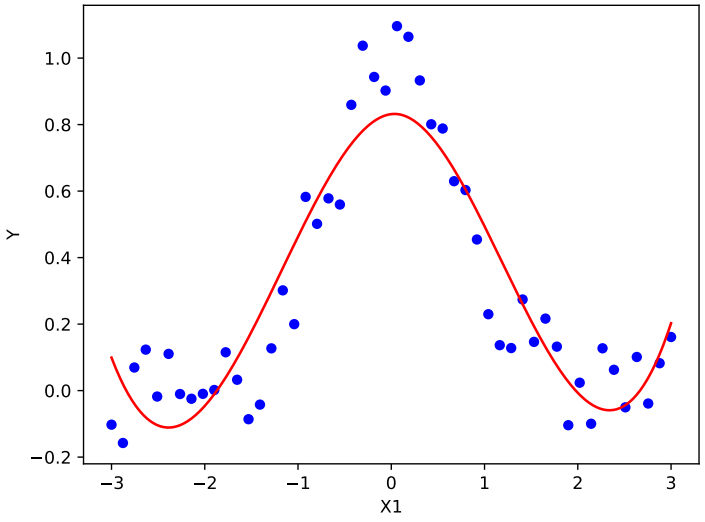

Figure 5: Smoothing spline with unknown parameter λ

(c) Suppose that a curve $\hat{g}$ is computed to smoothly fit a set of n points using the following formula:

$$\hat{g} = \arg\min_{g} \left(\sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int \left[g'''(x) \right]^2 dx \right) \quad .$$

To which value do we have to set λ to get a fit as shown in Figure 5 - i) $\lambda = 0$; ii) $\lambda = 1$; or, iii) $\lambda = \infty$? Explain your answer. (1 point)

Recall that a multiple linear regression model can be written as:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$$

Whereas a Generalized Additive Model (GAM) can be written as:

$$y_i = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip}) + \epsilon_i$$

(d) What are the functions $f_1 ... f_p$? Explain why they are introduced. (1 point)

(e) How does the logistic function for the Generalized Additive Model (GAM) formulation look like? Write it down. (1 point)

Solution.

(a) $K = 10$ hence $K + 1 = 11$ regions. $d = 3$ hence we fit a polynomial with $d + 1 = 4$ parameters. By only requiring the splines to be continuous up to the first derivative we introduce $K * 2$ constrains. Hence the degrees of freedom are

$$(K + 1)(d + 1) - 2K = Kd + d + 1 - K = 10 * 3 + 3 + 1 - 10 = 24$$

(b)

$$\begin{aligned} dof &= (K + 1)(d + 1) - cK \\ &= Kd + K + d + 1 - cK \end{aligned}$$

where c are the number of constrains per knot.

$$c = -\frac{dof - Kd - K - d - 1}{K}$$

$$c = -\frac{25 - 10 * 4 - 10 - 4 - 1}{10}$$

Hence $c = 3$, which means we have to enforce continuity up the the 2nd derivative.

(c) It cannot be i) because if $\lambda = 0$, there is no regularization effect and $\hat{g}$ would interpolate all points.
It cannot be iii) $\lambda = \infty$, since it would mean that the penalty term (integral) has to be 0. Function $\hat{g}$ would then have to be of degree 2, as this is the most complex function with a 3rd derivative equal to zero, but, judging by the number of local minima and maxima we can see that the function in Figure 5 has degree 4.
This leaves the only option: ii) $\lambda = 1$.
[See pg. 277, section 7.5.1 of ISLR]

(d) The functions $f_1, \ldots, f_p$ are smooth non-linear functions. They are introduced in order to allow for non-linear relationships between the feature and the response. This gives our model more expressive power than a linear model, which can only express a relationship with a scalar multiple of the feature value. This however, will not be sufficient in many cases. Consider an example of the relation between *age* and *income*. Intuitively, the expect that the income is low (or none) when one is an infant or young-adult (1 year to 18 years), increases thought the adult age (18 year to 60 years), then decreases as one retires (> 60 years). We can use non-linear functions of the *age* variable to express this, but cannot do the same with a linear function, which could only increase or decrease with age depending upon the sign of the coefficient.
[See pg. 283, section 7.7.1 of ISLR]

(e) Same as the standard logistic function for logistic regression, but now using the given equation as the exponent

$$p(x) = \frac{1}{1 + e^{-(=\beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip}) + \epsilon_i)}} .$$

Problem 5 (Unsupervised) **(10 points)**

Consider the following points that you wish to investigate for possible clusters:

i	X_1	X_2
1	1	5
2	1	4
3	2	0
4	-1	1
5	0	2

Table 2: Observations for predictor variable X_1 and X_2 for points $i = 1...5$.

(a) Can the following properties of the k-means clustering algorithm be seen as advantages or disadvantages? Explain why.
Properties: (1) algorithm convergence, (2) initialization procedure, (3) (in)sensitivity to outliers, (4) its (in)ability to cluster of new (unseen) samples. (2 points)

(b) Given the points in Table 3. Assume the k-means algorithm for $k = 2$ has not yet converged. At some step i you observe the means of cluster 1: $\hat{\mu}_1^{(i)} = (0, 4)$ and cluster 2: $\hat{\mu}_2^{(i)} = (1, 1)$. Perform the next step (one step) of the k-means algorithm. Report (i) the cluster assignments (i.e. which points belong to each of the three clusters) and (ii) the coordinates of the new cluster means (i.e. $\hat{\mu}_1^{(i+1)}$ and $\hat{\mu}_2^{(i+1)}$). Make your calculations explicit. (2 points)

(c) Consider Figure 6 (next page), where final cluster assignments and means are indicated using k-means clustering. Give an (informal) description of how would you find the decision boundaries between the clusters. (1 point)

(d) What is the difference between k-means and k-medoids? Given the cluster assignments in Figure 6, what would be the medoids of the respective clusters using the Euclidean distance as dissimilarity metric? (2 points)

(e) To perform k-means clustering and k-medoids, do we need to know the coordinates of the elements to be clustered, or does it suffice to only know their mutual distances? Explain why. (2 points)

(f) If you are interested in clustering the observed data in terms of the sign of their feature values (not the actual values) - would you still use Euclidean distance as dissimilarity metric? Explain why (not)? If not, what dissimilarity metric would you use for clustering? (1 point)

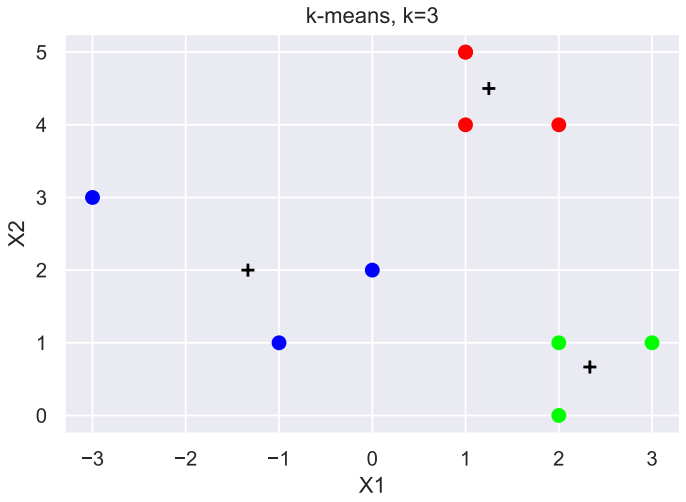

Figure 6: Converged k-means clustering with $k = 3$ using data from Table 3

Solution.

(a) Disadvantages:

2. Initialization is stochastic so several different initializations may be required to get a decent result.
3. Outliers can distort mean and pull it towards the outlier point. This can cause clusters to exclude points that would otherwise belong in them.

Advantages:

1. Guarantees convergence and provides a natural stopping criterion.
4. - Easily adapts to new examples. They can be assigned to the closest cluster mean.

(b) Euclidean distance $\sqrt{(x_1 - \hat{\mu_1})^2 + (x_2 - \hat{\mu}_2)^2}$.
Assign to cluster i with smaller $\delta^{(i)} = (x_1 - \hat{\mu}_1^{(i)})^2 + (x_2 - \hat{\mu}_2^{(i)})^2$ (no need to take square root).

i	X_1	X_2	$\delta^{(1)}$	$\delta^{(2)}$	cluster
1	1	5	2	16	1
2	1	4	1	9	1
3	2	0	20	2	2
4	-1	1	10	4	2
5	0	2	4	1	2

Table 3: Observations for predictor variable X_1 and X_2.

Cluster 1: { 1, 2 } with mean $\hat{\mu}_1^{(1)} = (0.5(1+1), 0.5(4+5)) = (1, 4.5)$
Cluster 2: { 3, 4, 5} with mean $\hat{\mu}_2^{(1)} = (0.33(2-1+0), 0.33(0+1+2)) = (0.33, 1)$

(c) The decision boundary between two clusters can be found by drawing a straight line between the two cluster means, then the decision boundary is orthogonal to it. When there are three clusters then, the decision boundaries will meet in the middle.

(d) Difference between k-means and k-medoids is that they differ in the computation of cluster centers. Means are computed with Eucledian distance, whereas Mediods are the actual data points that have the smallest distance to other points in their cluster. They can be computed given only the dissimilarity matrices.
Medoid of the red cluster: $(1, 4)$
Medoid of the blue cluster: $(-1, 1)$
Medoid of the green cluster: $(2, 1)$

(e) For k-means yes, as it needs a metric space to compute the centroids. It cannot be computed only on dissimilarity matrices. For k-medoids, however, it is enough to know just the mutual distances or dissimilarities to compute the medoid of each cluster.

(f) No, because sign does not play a role in Euclidean distances i.e. -2 is as far away from 0 as $+2$ is. We can construct a new metric by multiplying the signs of each feature. This metric would cluster observations sharing the same sign (be it positive or negative) in one cluster, having dissimilarity $+1$, and observations differing in sign in another cluster, with dissimilarity -1.