Elements of Machine Learning, WS 2023/2024
Krikamol Muandet and Jilles Vreeken
EXAM, FEBRUARY 23RD, 2024, SOLUTION SHEET

PROBLEM 1 (INTRODUCTION) **(10 points)**

1. Are the following statements correct or incorrect? Explain each answer. (7pts)

   (a) Among all classifiers presented in the lecture, Support Vector Machines always achieve the lowest test error, since it finds the maximum margin classifier. (1pt)

   (b) A feed-forward neural network without an activation function is equivalent to a linear model. (1pt)

   (c) The k-means algorithm is guaranteed to converge to the global optimum. (1pt)

   (d) The t-SNE embedding captures the directions of largest variance in the data. (1pt)

   (e) t-SNE is deterministic, i.e. for the same data it always gives the same result. (1pt)

   (f) k-fold cross-validation with $k < n - 1$ has higher bias than LOOCV. (1pt)

   (g) For data of a large enough sample size n, bootstrap sets that are sampled uniformly at random will be uncorrelated. (1pt)

2. Josephine is analyzing three datasets, but is not entirely happy with the results. Explain for each case how we can address her concerns. (3pts)

   (a) On Dataset 1 of $p = 50$ predictors for a real-valued outcome Y, Josephine applied polynomial regression ($d = 4$). She is concerned the model overfits and that some of the predictors are not relevant for predicting Y. (1pt)

   (b) On Dataset 2 of a single predictor X and real-valued outcome Y, Josephine fits a piecewise polynomial regression spline with $k = 5$ knots. When plotting the result, the fitted function looks discontinuous in the regions around the knots, making Josephine wonder how to make the model more smooth. (1pt)

   (c) On Dataset 3 of $n = 1000$ samples and $p = 100$ variables, Josephine applies k-means clustering to discover k clusters. How can she interpret the results, for example, to verify if the clusters make sense? (1pt)

Solution.

1. (a) No. Counterexample: XOR problem. (1pts)

   (b) Yes. A feed-forward neural network with a linear activation function can be re-written as a linear model with coefficients $\beta = \prod_i W_i$, where W_i denotes the weights of the respective layer. (1pts)

   (c) No. K-means is only guaranteed to converge to a local optimum. (1pt)

   (d) No. This is the objective of PCA, not t-SNE. (1pt)

   (e) No. t-SNE is a stochastic algorithm. (1pt)

   (f) Yes. Higher bias of the learned model cause we hold out larger validation set. (1pt)

   (g) No. Bootstrapping samples with replacement and the resulting sets will always be correlated. (1pt)

2. (a) Adding L1 regularization, like in Lasso, can mitigate overfitting and allow for variable selection. Alternatively, we can address overfitting by reducing the degree d, doing Ridge regression, and do subset selection, cross validation or similar to select relevant variables.

   (b) Adding continuity constraints on the derivatives as seen in the lecture.

   (c) Dimensionality reduction technique such as t-SNE.

Problem 2 (Linear Regression) (10 points)

1. Data scientists Ali Prediktørson and Omer Régrèssionaire investigate the effects of two predictors X_1, X_2 on a real-valued outcome Y. Using a linear regression model $\hat{f}$, they obtain the following coefficients and standard errors (intercept not shown),

	coefficient	std. error
	β	$SE(\beta)$
X_1	3.31	0.04
X_2	-0.28	0.0012

   (a) What can you say about the relationship between X_1, X_2 and Y based on the coefficients β? (1 pt)

   (b) State the 95% confidence interval for the coefficient of X_1. (1 pt)

   (c) The dataset has $n = 103$ samples, a $TSS = 1500$, and the above model has an $RSS = 600$. Explain step by step how to design a hypothesis test to decide whether *at least one* of the variables X_1, X_2 is relevant for predicting Y. (1 pt)

2. Omer plots the residuals of their model for X_1 in Figure 1.

   (a) Ali argues that the points x_2 and x_3 have unusual values of X_1 compared to the population and suggests removing them from the dataset to fit a more reliable model. Do you agree? Why? (1 pt)

   (b) Ali suggests to do Leave-One-Out Cross-Validation (LOOCV) three times, each time leaving out one of the points x_1, x_2, and x_3. Based on the residual plot, which of these do you expect to result in the highest train, respectively test error? Why? (1 pt)

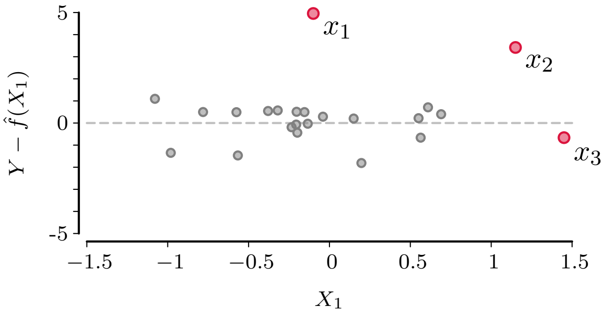
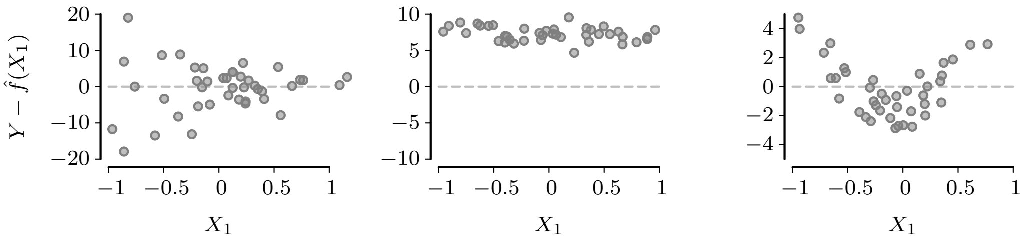


Figure 1: Residual Plot for Problem 2.2



(i) Residuals for D_1.

(ii) Residuals for D_2.

(iii) Residuals for D_3.

Figure 2: Residual Plots obtained by Filippo for Problem 2.4

3. Filippo Speçulatio is convinced that we should include another predictor X_3. He makes the following claims. For each, determine whether it is true or false, and give a brief explanation why. (2 pts)

   (i) The R^2 score of the model that includes X_3 in addition to X_1, X_2 will always be larger than that over the one with only X_1, X_2.

   (ii) As the predictor X_1 is significant in the model with only X_1, X_2, it will still be significant when we include X_3.

   (iii) We can check which predictors are useful by adding a regularization term.

   (iv) We can check whether the RSS values for any two predictors are correlated with each other to find out whether there is an interaction between two predictors. If so, we should add an interaction term to the model.

4. Filippo applies the linear model from Problem 2.1 to three different datasets, D_1, D_2, and D_3. He forgets to fit the intercept term. He plots the residual plots in Figure 2. (3 pts)

   (a) For each of the three datasets, explain why a linear model is suited or not. ($1\frac{1}{2}$ pts)

   (b) For each of the three datasets, propose an appropriate change to the model to improve the prediction accuracy. Explain your reasoning. ($1\frac{1}{2}$ pts)

Solution.

1. (a) Positive correlation with the outcome for X_1, and negative correlation with the outcome for X_2.

   (b) Confidence interval $[\beta - 2\text{SE}(\beta_1), \beta_1 + 2\text{SE}(\beta_1)] = [3.23, 3.39]$.

   (c) Hypothesis testing using $F-$statistic

   $$\frac{\text{TSS} - \text{RSS}}{\text{RSS}} \frac{n-p-1}{p}$$

   Here, resulting in F-statistic 150. The corresponding p-value can be used for rejecting the null-hypothesis $H_0 : \beta_1 = \beta_2 = 0$. (Students don't need to compute values but list the steps.)

2. (i) True, R^2 monotonically increases even if we add irrelevant predictors.

   (ii) False, X_1 may be insignificant in the joint model, for example, if it's corellated with X_3.

   (iii) True, Lasso constrains the coefficients of irrelevant predictors to zero.

   (iv) False, this indicates that the predictors are corellated rather than interacting.

3. (a) Among the three options, leaving out x_1 or x_2 will improve train error the most, for x_3 little effect. Conversely, small test error for x_3 but large for the others.

   (b) No, better to remove the outliers x_1, x_2.

4. (a) D_1: heteroskedastic, can modify model to weighted linear regression.

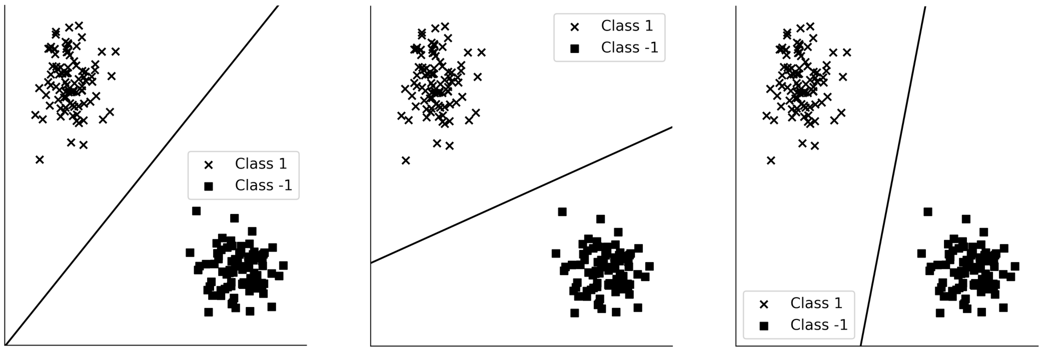
   (b) D_2: linear, should fit intercept.

   (c) D_3: polynomial (here $d = 2$), should use nonlinear regression method.

PROBLEM 3 (CLASSIFICATION) **(10 points)**

1. (a) Give the decision boundary of a Maximal Margin Classifier as a function $g(x) = 0$, $x \in \mathbb{R}^p$. What is the geometric interpretation? (1 pt)

   (b) Consider Figure 3. Which of the three plots correspond to the decision boundary of a Soft-Margin Support Vector Classifier with $C = 0$? Explain your answer. (1 pt)



(i) (ii) (iii)

Figure 3: Three decision boundaries

   (c) What is the primary limitation of a hard-margin SVM and how does a soft-margin SVM resolve it? Sketch a dataset that showcases the problem. (2 pt)

2. (a) What problem does Christina face when she would apply Linear Discriminant Analysis (LDA) on the data depicted in Figure 4? Which classifier would you recommend her instead? Why? (1 pt)

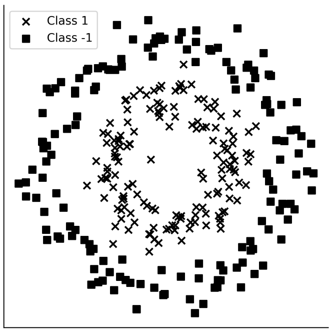


Figure 4: Plot belonging to Problem 3.2.

   (b) In the lecture, we derived LDA using Bayes' rule. Starting from (3 pt)

$$p_k(x) = \Pr(Y = k \mid X = x) = \frac{\pi_k f_k(x)}{\sum_{\ell=1}^{K} \pi_\ell f_\ell(x)} ,$$

   we assumed that $f_k(x)$ is a univariate Gaussian with the same variance across all classes.

Christina proposes that we should assume that each class follows a Rayleigh distribution. The density of the Rayleigh distribution is given by

$$f_k(x) = \begin{cases} \frac{x}{\sigma_k^2} e^{\frac{-x^2}{2\sigma_k^2}} & , x \geq 0 \\ 0 & , x < 0 \end{cases} \quad \text{, where } \sigma_k > 0.$$

Derive the discriminant **and** the decision boundary for $x \geq 0$. Is the discriminant linear in x? You can assume that the parameters are chosen such that we do not divide by 0.

3. Muhammed claims that Logistic Regression is a non-linear model and shows Figure 5 as evidence. Is he right? Explain your answer. (2 pt)

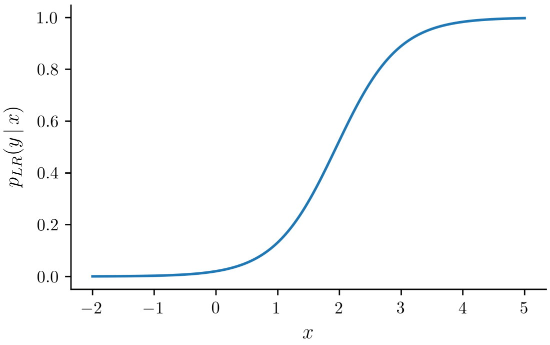


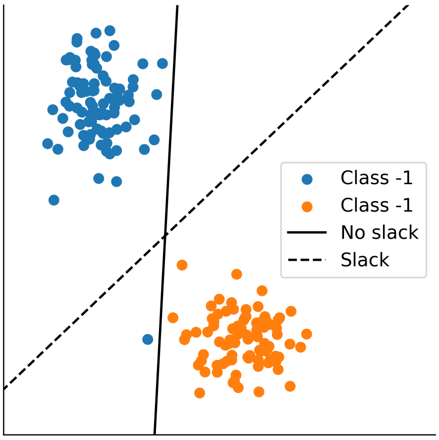
Figure 5: Plot for Problem 3.3

Solution.

1. (a) $g(x) = w^T x + b = 0$, where w is the weight vector, b is the bias term, x the input. The decision boundary is a hyperplane in $\mathbb{R}^p$.

   (b) Figure (a) corresponds to the SVM decision boundary, as it has the biggest margin to both classes. The DB in Figure (b) and (c) either closer to class -1 or class 1.

   (c) A outlier or wrongly labeled datapoint results in bad decision boundary for the Hard-margin SVM.



2. (a) The dataset is not linearly seperable, hence LDA can not find a hyperplane to seperate the classes. Use an SVM with an RBF kernel.

   (b)
   - Discriminant: $\delta_k(x) = \ln(\pi_k f_k(x)) = \ln(\pi_k) + \ln(x) - \ln(\sigma_k^2) - \frac{x^2}{2\sigma_k^2}$
   - Decision boundary: In the following we assume that parameters are chosen such that we do not divide by 0.

$$\begin{aligned} \delta_k(x) - \delta_l(x) &= 0 \\ \ln(\pi_k) - \ln(\sigma_k^2) - \frac{x^2}{2\sigma_k^2} - \left(\ln(\pi_l) - \ln(\sigma_l^2) - \frac{x^2}{2\sigma_l^2}\right) &= 0 \\ x^2\left(\frac{1}{2\sigma_l^2} - \frac{1}{2\sigma_k^2}\right) &= \ln(\pi_l) - \ln(\pi_k) - \ln(\sigma_l^2) + \ln(\sigma_k^2) \\ x^2 = \left(\ln(\pi_l) - \ln(\pi_k) - \ln(\sigma_l^2) + \ln(\sigma_k^2)\right) &\left(\frac{1}{2\sigma_l^2} - \frac{1}{2\sigma_k^2}\right)^{-1} \end{aligned}$$

   - Discriminant is non-linear

3. Logistic regression is considered a linear model because it makes use of a linear combination of input features. The term "linear" in this context refers to the linearity of the relationship between the input features and the log-odds (logit) of the output.

The logistic regression model takes the form:

$$P(y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n)}}$$

Problem 4 (Unsupervised) **(10 points)**

1. Jawad and Ahmed argue about what is the best laptop. Ahmed claims his is better because it has the new M4 processor, Jawad disagrees and says his is better because it has a touch display.

   As they are data scientists, they agree to settle the matter by examining a dataset $X \in \mathbb{R}^{n \times p}$ of $n = 10\,000$ laptops for which they collect $p = 1\,000$ different features.

   They first want to inspect the data visually. For this, they use PCA to reduce their dataset to $p = 2$ dimensions.

   (a) Briefly describe the general idea behind PCA. (1 pt)

   (b) We know that the principal components of X are the eigenvectors of the covariance matrix $X^T X$. Show that the unit eigenvector w_1 with the largest eigenvalue λ_1, where it holds that (3 pts)

   $$X^T X w_1 = \lambda_1 w_1 \ , \lambda_1 > \lambda_2 > \cdots > \lambda_p,$$

   is the principal component that maximizes the variance of the projected data.

   (c) Another view on PCA is that it tries to minimize the reconstruction error of the data. Explain how reconstruct the projected data back to the original space and what error the first principal component obtains in comparison to the other principal components. (1 pt)

2. After applying PCA, they use k-means to cluster the data into 5 clusters. (1 pt)

   Jawad runs K-means on his Microsoft Surface Pro and gets clustering with an in-cluster variation of 1000. Ahmed runs k-means on his M4 Macbook Pro and gets a clustering with an in-cluster variation of 500. Ahmed argues that his result is better because he is using Apple Silicon. Describe another reason why they could be getting different results that is not related to the processor.

3. Consider the following two initialization strategies for k-means:

   - Randomly pick K points from the dataset as the initial centroids.
   - Randomly assign each point to a cluster and compute the initial centroids as the averages of these clusters.

   (a) For each of the two strategies, where are the initial centroids expected to be on average, and how much variance would the centroids have between different initializations? (2 pts)

   (b) For each of the two strategies, describe an advantage over the other, when we employ them in standard k-means. (2 pts)

Solution.

1. (a) The general idea behind PCA is to find a lower dimensional representation of the data that captures the most variance. The data is linearly projected onto a lower dimensional subspace, while trying to minimize the reconstruction error.

   (b) The variance of the data projected onto a principal component w is given by

$$\frac{1}{n}\sum_{i=1}^{n}\left(w^T x_i\right)^2 = \frac{1}{n}w^T\left(\sum_{i=1}^{n} x_i x_i^T\right)w = \frac{1}{n}w^T X^T X w$$

   As w is an eigenvector of we know that

$$\frac{1}{n}w^T X^T X w = \frac{1}{n}\lambda w^T w \;.$$

   As $||w|| = 1$, we have that the variance is given by $\frac{\lambda}{n}$. This term is maximized by λ_1 as it is the largest eigenvalue of $X^T X$.

   (c) To reconstruct the projected data, we simply multiply the projected data with the principal components and add the mean of the original data if it was subtracted before, i.e.

$$\hat{X} = XWW^T + \bar{X} \;.$$

   The first principal component minimizes the reconstruction error, as it maximizes the variance of the projected data.

2. The initialization of K-means is random, so it is possible that Jawad's clustering is worse than Ahmed's clustering simply because of the initialization.

3. (a)
   - If we randomly pick K points as initial centroids, then the centroids are spread uniformly throughout the entire domain, with each point having a probability of $1/n$ to be included. Therefore, the expected value of the initial centroids is the mean of the data $\bar{x}$, and the variance is given by $Var(X)$.
   - For the second approach, we first randomly assign each point to a cluster. Still, the initial centroid is expected to be $\bar{x}$, but the variance is smaller. This is because we use a subsample of the data to compute the initial centroids, which reduces the variance.

   (b)
   - Using the first strategy, we expect to have more spread out initial centroids, which is beneficial for K-means to find distinct clusters. This is a downside of the second strategy, as the initial centroids are more likely to be close to each other.
   - A downside of the first strategy is that outliers could be chosen as initial centroids, which could lead to bad results, near empty clusters due to initializations. However, this is less likely to happen with the second strategy, as the initial centroids are computed using a subsample of the data where outliers are easily averaged out.

Problem 5 (Trees and Splines) **(10 points)**

1. You are given a dataset with points from three different classes and want to classify them based on a decision tree. The plot below illustrates the data points (class labels are indicated by the symbols $[\times, \triangle, \circ]$) and the decision boundaries of a decision tree.

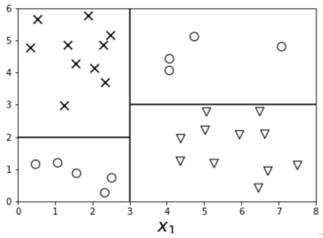


Figure 6: Decision tree Plot belonging to Problem 5.1.

   (a) Draw the corresponding decision tree. Make sure that you include the features (X_1 or X_2) and thresholds of the split. For each node in the tree, also give the number of training samples per class that arrive in that node. (1 pt)

   (b) Compute the Gini index for all internal nodes and all leaves in your decision tree. *Note: Your answer may contain improper fractions (e.g. $\frac{117}{33}$).* (2 pt)

2. (a) Assume we have a dataset of two predictive variables X_1 and X_2, with two different classes C_1 and C_2. The points from class C_1 are given by $A = \{(i, i^2) \mid i \in \{1 \dots 100\}\} \subseteq \mathbb{R}^2$, while the points from class C_2 are $B = \{(i, \frac{125}{i}) \mid i \in \{1 \dots 100\}\} \subseteq \mathbb{R}^2$. Construct a decision tree of minimal depth that assigns as many data points as possible to the correct class. Provide for each split the feature and corresponding thresholds. How many and which data points are misclassified? (2 pt)

   (b) Assume we have a dataset D of n samples (x_i, y_i) with $x_i \in \mathbb{R}^2$ and $y_i \in \{0, 1\}$. We aim to train a decision tree using entropy as the splitting criterion. We stop building the tree when there is zero *improvement* in purity for all splits. (1 pt)
   Give an example of a small dataset D that contains at least one instance from each class, and for which the learned decision tree has no splits – the root node is a leaf. Justify your answer.
   Hint: you do not need more than a few instances.

3. Assume a linear spline of K knots, i.e. $f(x_i) = \beta_0 + \beta_1 x_i + \sum_{k=1}^{K} b_k (x_i - \xi_k)_+$ where $(x_i - \xi_k)_+ = \max(x_i - \xi_k, 0)$ and b_k are the spline coefficients. We aim to minimize the sum of squared residuals

$$\text{minimize} \quad S = \sum_{i=1}^{N} ||y_i - f(x_i)||_2^2 .$$

   (a) What is the degree of freedom for this model? Explain your answer. (1 pt)

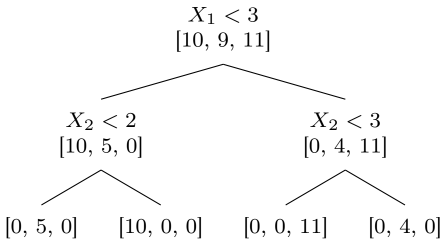
(b) To avoid overfitting, we usually introduce a penalty on the spline coefficients such as $\sum_{k=1}^{K} b_k^2$. Therefore we minimize a modified objective with regularization parameter λ, (3 pt)

$$\text{minimize} \quad S + \lambda \sum_{k=1}^{K} b_k^2 \, .$$

Derive the closed-form solution of the optimal parameters $\hat{\beta} = [\beta_0, \beta_1 \ldots, \beta_d, b_1 \ldots, b_k]^T$

Solution.

1. (a)



(b)

$$\text{Gini index: } i_G(t) = \sum_c p_i(1-p_i) = 1 - \sum_c p_i^2$$

$$\text{Root node: } i_G(t) = 1 - \left(\frac{10}{30}\right)^2 - \left(\frac{9}{30}\right)^2 - \left(\frac{11}{30}\right)^2 \approx 0.664$$

$$\text{Left child of root: } i_G(t) = 1 - \left(\frac{10}{15}\right)^2 - \left(\frac{5}{15}\right)^2 - (0)^2 \approx 0.444$$

$$\text{Right child of root: } i_G(t) = 1 - \left(\frac{0}{15}\right)^2 - \left(\frac{4}{15}\right)^2 - \left(\frac{11}{15}\right)^2 \approx 0.391$$

$$\text{Left leaf node: } i_G(t) = 1 - \left(\frac{5}{5}\right)^2 - \left(\frac{0}{5}\right)^2 - \left(\frac{0}{5}\right)^2 = 0$$

$$\text{Left-middle leaf node: } i_G(t) = 1 - \left(\frac{10}{10}\right)^2 - \left(\frac{0}{10}\right)^2 - \left(\frac{0}{10}\right)^2 = 0$$

$$\text{Right-middle leaf node: } i_G(t) = 1 - \left(\frac{0}{10}\right)^2 - \left(\frac{0}{10}\right)^2 - \left(\frac{10}{10}\right)^2 = 0$$

$$\text{Right leaf node: } i_G(t) = 1 - \left(\frac{0}{4}\right)^2 - \left(\frac{4}{4}\right)^2 - \left(\frac{0}{4}\right)^2 = 0 \text{ (all leaf nodes)}$$

2. (a)
   - Split root node based on feature 1 < 5
   - Split both child nodes based on feature 2 ≤ 25
   - $(5, 25)$ is in both classes and misclassified.

   (b) One possible solution is $\{(1,1),1\}, \{(-1,-1),1\}, \{(-1,1),0\}, \{(1,-1),0\}$

3. (a) The degree of freedom is $K+2$ since there are $K+2$ free parameters.

   (b) We aim to minimize the sum of squared residuals with a penalty on the size of the coefficients. The objective function is:

$$\text{minimize} \quad S = \sum_{i=1}^{N} \left(y_i - \beta_0 - \beta_1 x_i - \sum_{k=1}^{K} b_k (x_i - \xi_k)_+ \right)^2$$

where $(x_i - \xi_k)_+ = \max(x_i - \xi_k, 0)$ and b_k are the spline coefficients. The penalty term is added to the objective function as follows:

$$\text{minimize} \quad S + \lambda \sum_{k=1}^{K} b_k^2$$

We can write the objective function in matrix notation:

$$\text{minimize} \quad (y - X\beta)^T(y - X\beta) + \lambda\beta^T D\beta$$

Differentiating with respect to β and setting the derivative to zero gives us:

$$\frac{\partial}{\partial\beta}\left[(y - X\beta)^T(y - X\beta) + \lambda\beta^T D\beta\right] = 0$$

Expanding and simplifying, we have:

$$-2X^T(y - X\beta) + 2\lambda D\beta = 0$$

Rearranging terms, we arrive at the normal equations for the penalized least squares problem:

$$(X^T X + \lambda D)\beta = X^T y$$

Solving for β, we obtain the estimate:

$$\hat{\beta} = (X^T X + \lambda D)^{-1} X^T y$$

where D is a diagonal matrix with the first two diagonal terms zero (no penalty on β_0 and β_1) and the rest are 1 (penalty on the b_k coefficients).