Question Booklet

- **The final exam contains 5 questions and is scheduled for 2.5 hours. At maximum you can earn 50 points.**
- **Please verify if this question booklet consists of 10 pages, and that all questions are readable, else contact the examiners immediately.**
- **This is an open-book exam. You are allowed to consult the books, slides, and lectures while writing it. You are not allowed to consult others. Plagiarism is not condoned.**
- **Answers without sufficient details are void: you get no points for "yes" or "no" answers, unless the question specifically asks this. Explain all answers in your own words. Clearly state any assumptions you make.**
- **No cats, interns, or other pets were harmed when preparing this exam.**

Problem 1 (Errors, Errors Everywhere) **(10 points)**

(a) Consider Figure 1. Which of the following three options correctly describes what is happening in the figure. (1 point)

i) As we increase flexibility, variance starts to increase. The bias decreases more rapidly than the increase in variance, hence causing a downward trend until *Flexibility* = 6. After that, the decrease in bias becomes smaller than the increase in variance, resulting in the upward trend in the curve.

ii) As we increase flexibility, bias starts to increase. The variance decreases more rapidly than the increase bias, hence causing a downward trend until *Flexibility* = 6. After that, the decrease in variance becomes smaller than the increase in bias, resulting in the upward trend in the curve.

iii) As we increase flexibility, both bias and variance decrease until *Flexibility* = 6. After that, bias approaches zero, but variance starts to increase due to over fitting, thereby causing the overall rising trend in the curve.

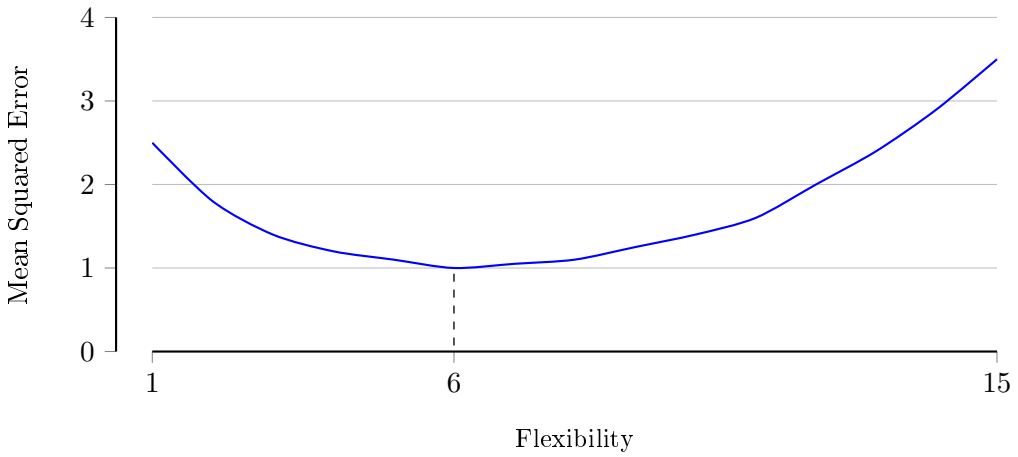


Figure 1: Test MSE for an un unknown data set.

(b) Explain, for each of the three settings below, what will happen in terms of bias and variance when we make the proposed change to the learning procedure. (3 points)

1) Changing the maximum depth of a decision tree from 10 to 2,

2) Replacing the LDA classifier with the QDA classifier,

3) When fitting a hexa-spline (i.e. polynomial spline of degree 6) enforcing continuity up to the 2nd, instead of up to the 5th derivative.

(c) Consider that we have infinite patience and training data. Is it *in general* possible to achieve the perfect predictor that obtains 0 test error? Explain why (not). (1 point)

(d) Rank the following approaches from the one that over-estimates generalization error *least* to the one that over-estimates generalization error *most*. Explain why your first ranked method over-estimates less than the second ranked method, the second ranked method less than the third ranked method, and so on. (3 points)

- validation set,
- leave-one-out cross-validation (LOOCV),
- k-fold cross-validation (CV),
- bagging.

(e) Explain *in your own words* what a confidence interval is, what a prediction interval is, and how these two are different. (2 points)

Problem 2 (Regression) **(10 points)**

Please assist a group of experts with analyzing the 5 data points they obtained through a highly expensive experiment involving deep quantum entanglement and other buzzwords. The key objective is to predict response variable Y given one or more predictors X.

Expert 1 is convinced that X_1 is the key to predicting Y, and asks you to analyze the data in Table 1 using linear regression.

X_1	Y
2.3	15
2.7	14
3.8	16
3.9	15
4.6	24

Table 1: Observations for predictor variable X_1 and target variable 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 $\mathbf{X} = [\mathbf{X_1}; \mathbb{1}]$.

(a) Using the following convenient approximation,

$$(\mathbf{X^T}\,\mathbf{X})^{-1} = \begin{bmatrix}0.3 & -1 \\ -1 & 3.6\end{bmatrix},$$

find β_1 and β_0. Explain the *reasoning* behind each step. (3 points)

(b) Expert 1 insists on using an unbiased linear estimator. Using the exact $(\mathbf{X^T}\,\mathbf{X})^{-1}$ under which conditions can you guarantee that the estimated β_0 and β_1 will have the smallest Standard Error? (1 point)

Expert 2 does not like X_1 at all, and instead claims that predictor X_2 is linearly related with the response variable. As evidence they show you the plot given in Figure 2 on page 5.

(c) State the common name of this plot, and describe what it is useful for. (1 point)

(d) Is Expert 2 correct in their claim? Explain why (not). (1 point)

Expert 3 is more inclusive and considers it possible that not just X_1, or X_2, but rather that any or all of $X_1, X_2, \ldots, X_{42}$ are useful for predicting Y. They kindly provide data (not shown) over all 42 predictors for the same observations as given in Table 1.

(e) Describe why you can no longer use the same general approach to determine the linear regression coefficients as you could to help Expert 1. (1 point)

Finally, the head of research unit, Expert ∞, wants to know which from $X_1, \ldots, X_{42}$ are the most relevant predictors for Y and suggests that to find out you should use ridge regression. An intern points out that this approach may have its pitfalls.

(f) Give one reason in favor of using lasso over ridge regression, and another reason why to favor ridge regression over lasso. (2 points)

(g) Both ridge regression and the lasso require you to choose a value for λ. Describe how in this case you would choose a suitable value for this parameter. (1 point)

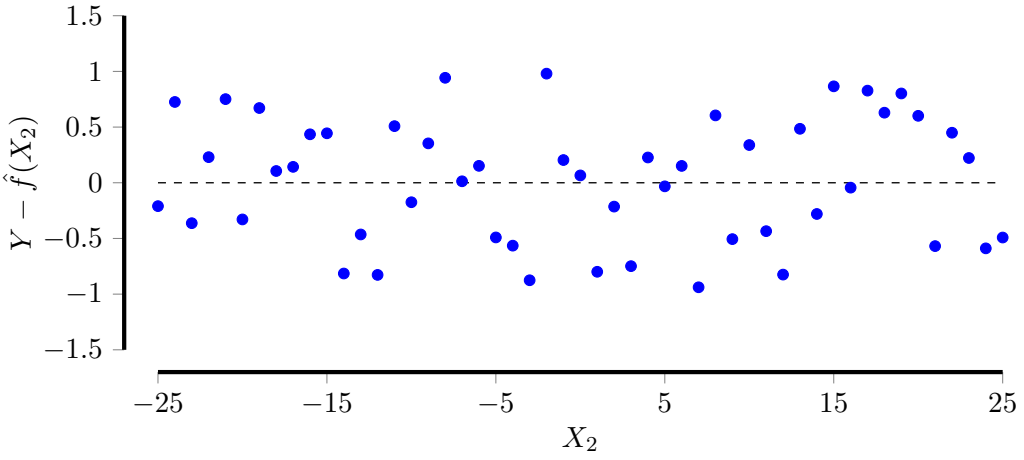


Figure 2: Plot given to you by Expert 2

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

The research team now considers a classification problem in which they want to predict whether a cat in a box is alive ($Y = 1$) or not ($Y = 0$).

Expert 1 advocates they should use Decision Trees, as these permit interpretable models.

(a) When growing a decision tree, we iteratively split the n data points of a node t over two successor nodes t'_1 and t'_2. Show that the classification error of a decision tree never increases when we do so. For simplicity, you may consider the class label to be binary, i.e., $p(c_0) + p(c_1) = 1$. (2 points)

Expert 2 mumbles something about variance, and argues to use a Support Vector Machine instead. Recall that the Support Vector Machine is defined as follows.

$$\begin{aligned} \underset{\beta_0,\ldots,\beta_p,\ \xi_1,\ldots,\xi_N}{\text{maximize}} \quad & M \\ \text{subject to} \quad & \|\beta\| = 1 \\ & \xi_i \geq 0 \\ & y_i f(x_i) \geq M(1 - \xi_i) \quad \text{for } i = 1, \ldots, N \\ & \sum_{i=1}^{N} \xi_i \leq C \end{aligned}$$

(b) Explain the purpose of variable C. Describe how bias and variance change when C is increased. (2 points)

Expert ∞ says they should not waste time with tweaking parameters and instead immediately go for the Bayes Classifier because it is *ideal*. The intern looks panicked.

(c) Give the Bayes Classifier. Explain in what sense it is ideal, and why we do not use it in practice so often. (1 point)

Expert 3 remarks that cats have nine lives, and that the classification problem hence involves not two, but rather $K = 10$ 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*,

$$\Pr(Y = k \mid X = x) \ \propto \ \pi_k \cdot f_k(x) \ .$$

(d) What is π_k in the above equation? How would you calculate π_k for a given data set? (1 point)

(e) Assume that f_k is provided. How can you now use Bayes' theorem to predict the class label given the training data. (1 Point)

The intern figured that x is univariate and always positive. Moreover, they strongly suspect that it follows an exponential distribution $\mathcal{E}(\lambda_k)$ with distinct λ_k for each of the k classes, where

$$\mathcal{E}(x; \lambda_k) = \lambda_k \cdot e^{-\lambda_k x} \ .$$

(f) Derive the discriminant function. (2 points)

(g) Is the above derived discriminant function linear in terms of x? Why (not)? (1 point)

Problem 4 (Beyond Linear Regression) **(10 points)**

The cat escaped from the box. The research team is now investigating how the size of a pet (X_1) relates to how loud it is (Y) as measured in decibel. Prior analysis shows that this relationship is non-linear.

Expert 1 suggests to model the relationship between X_1 and Y using regression splines, as these allow us to easily check and/or control the degrees of freedom of the model.

(a) How many degrees of freedom does a regression spline have if we use polynomials of degree $d = 4$, have $K = 10$ knots, and require the spline to be continuous at the knots, but do not care about the continuity of the derivatives. Explain your answer. (1 point)

Expert ∞ proclaims that to improve generalization they should use *unnatural* cubic splines. These are plain cubic splines with polynomials of degree $d = 10$ at the boundaries, where at each knot we enforce continuity up to and including the second derivative.

(b) How many degrees of freedom has an unnatural cubic spline with $K = 10$ knots? (1 point)

(c) Will the unnatural cubic regression spline achieve better generalization than a regular cubic spline? Explain why (not)? (2 points)

Expert 2 suggests they should use local regression using a uniform weight function (kernel) over the k points closest to the query point x_0.

(d) In the worst case, how many different local models would we have to fit if we have n training points and are asked to make m independent predictions? (2 points)

Expert 3 tells the intern that PCA is just linear regression, and that PLS and gradient boosting have nothing to do with one another. The intern looks doubtful.

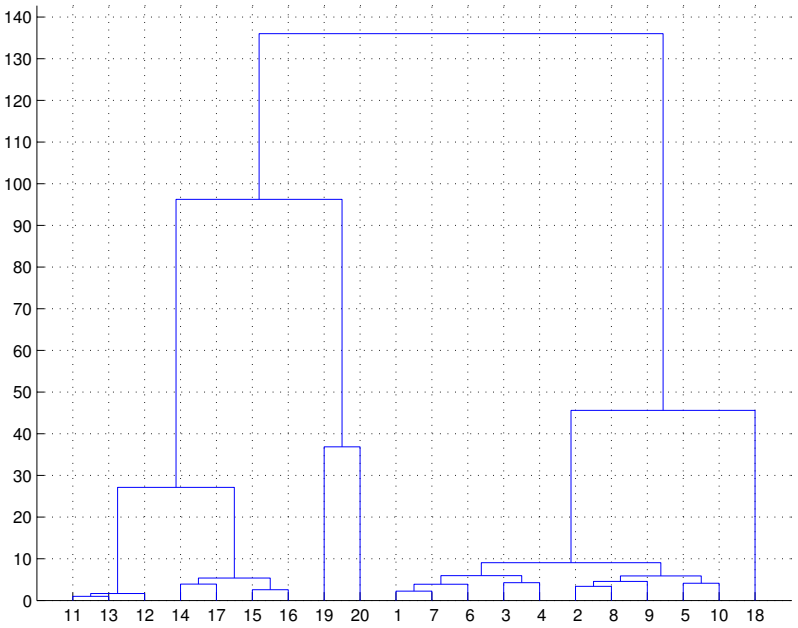
(e) Suppose we are given a zero centered dataset over X and Y with $\mathrm{Var}(X) = \mathrm{Var}(Y) = 1$. We fit a linear regression model from X to Y to obtain β_0 and β_1, respectively perform PCA over X and Y to obtain the first principle component Z_1. When we now compare the directions of the vector $(1, \beta_1)$ to that of Z_1, we see that these directions are similar yet different. Explain why. (2 points)

(f) Explain how PLS and gradient boosted regression trees are similar. What is an advantage of PLS over boosting and advantage of boosting over PLS. (2 points)

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

Having tried and solved all possible supervised machine learning problems in their field, the experts now want to gain *insight* from their data and hence turn to unsupervised learning.

Expert 1 considers hierarchical clustering. Consider the following dendrogram.



(a) What is the clustering at $k = 3$? (1 point)

(b) Expert 1 expects that the data has 3 real clusters, and a number of possible outliers. What are the clusters and outliers? Explain your choice. (2 points)

(c) Explain the difference between single-link and complete-link hierarchical agglomerative clustering. Which weaknesses of single-link does complete-link address? (2 points)

Expert 2 does not like hierarchies, and hence instead considers k-means clustering.

(d) Show why the k-means algorithm always converges. (2 point)

(e) Is k-means is sensitive to outliers in the data? Explain why (not). (1 point)

Expert ∞ likes looking at things. While Stochastic Neighborhood Embedding (SNE) has a neat formal definition, its results tend to suffer from the 'crowding' problem. The intern suggests that changing the distribution in the lower dimensional space might solve that. Expert ∞ tells the intern to use the probability density function sketched in Figure 3.

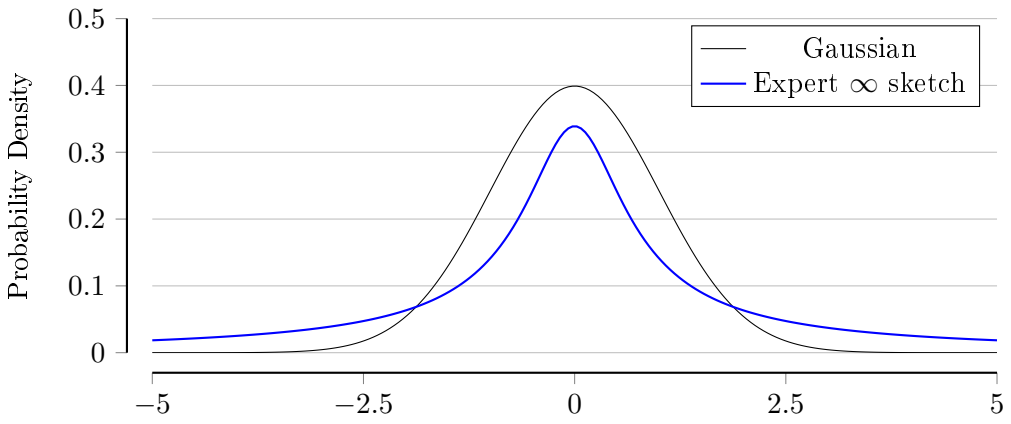


Figure 3: Probability Density Functions

(f) Consider Figure 3. Explain how and why the embeddings SNE discovers would change if we replace the Gaussian (Normal) distribution in the lower dimensional (map) space, with the sketched function. (2 points)