## Problem 1 (Linear Regression) (10 points)

1. Determine if the following statements are true or false. For every false statement, either correct it by replacing a *single* term (noun or adjective) *or* provide a counter example. (3pts)

   (i) Consider a linear model that consists of 3 predictors. We typically use the t-test to measure the collinearity of a single predictor with any of the rest.

   (ii) For a linear model that satisfies the least square assumptions, the $R^2$ statistic follows a $t$-distribution.

   (iii) Removing a high-leverage point always increases the accuracy of a linear model estimated using least squares.

   (iv) The least squares estimator can always be computed.

2. Not-yet-famous researcher Kanis Jalofolias is super interested in how the size ($X_s$) and weight ($X_w$) of a cat affects the loudness of its meow ($Y$). Each day he collects a dataset of all meows he encounters on random stray cats, normalises the predictors and creates a dataset. Since he knows that the relation must fulfil the OLS assumptions, he used this method on each dataset to fit a model (2pts)

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_w X_w + \hat{\beta}_s X_s \,.$$

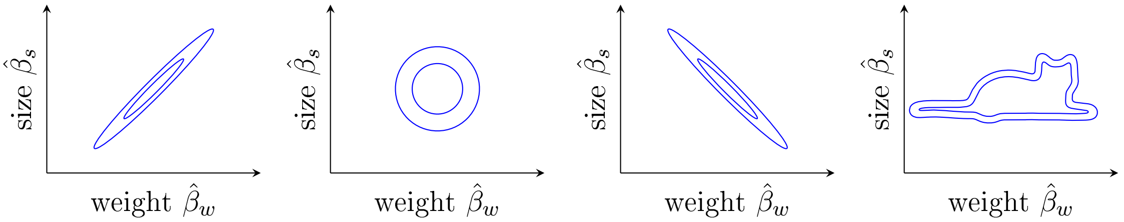
   After a few days he has a collection of models, and he creates a scatter plot where each point corresponds to the coefficients of each predictor for a given model. Which of the scatter plots below corresponds to the hypothetical case where:

   (i) The predictors *weight* and *size* are uncorrelated?

   (ii) The predictors *weight* and *size* are very positively correlated?

   (iii) The predictors *weight* and *size* are very negatively correlated?

   You are given 4 plots out of which you need to only use 3. Briefly explain all your choices and also why you left out the plot without a pair.



(a) Case A (b) Case B (c) Case C (d) Case D

Figure 1: Contour plots of the joint probability density of coefficient estimates $\mathbb{P}(\hat{\beta}_w, \hat{\beta}_s)$.

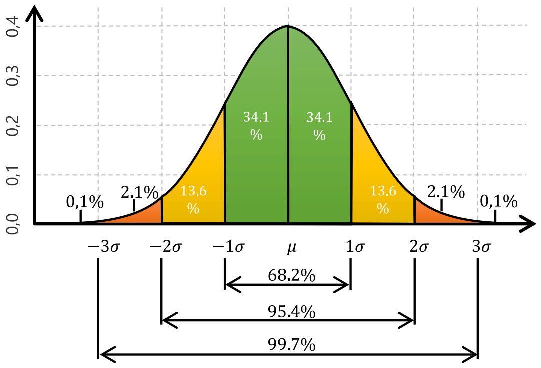
3. NopeAI recently received a large investment, and now wants to predict the added revenue based on two predictors $X_{\mathrm{bf}}$ and $X_{\mathrm{tw}}$, corresponding to its advertising budget for two different media, Bacefook and Twutter.

   (a) How can we verify whether a linear model is the right choice? (1pt)

   (b) Give an example scenario where adding an interaction term between Bacefook and Twutter would lead to a better model. (1pt)

   (c) World-famous CEO of NopeAI, Melon Usk, says that the prediction accuracy will improve if we add predictor $X_{\mathrm{bf}}^{99}$ to the model. How can we determine if this improvement is significant? (1pt)

4. Not-yet-famous researcher Kavid Daltenpoth considers a linear regression task with only one predictor. Using OLS he obtains an estimated $\hat{\beta}$. The variance of the estimate is $\mathrm{Var}(\hat{\beta}) = 0.49$. He looks at the following plot



and concludes:

> "*If X was uncorrelated with the outcome, with a probability* 2.2% *the value of* $\hat{\beta}$ *computed on a random dataset drawn from the true model would be greater than the one I just found.*"

   (a) Write down the equation that describes the fact Kavid stated. What is the common name for this probability? (1pt)

   (b) What is the value of $\hat{\beta}$? (1pt)

*Solution.*

1. (a) False
   We typically use the VIF for this.

   (b) False
   It follows an $F$-distribution.

   (c) False
   A simple counter-example is a high-leverage point that lies on the regression line. In this case, removing this point will very likely decrease the accuracy of the model.

   (d) False
   We can only compute a unique least square estimator when the covariance matrix is full rank, so that it can be inverted. A simple counter example is when the number of samples $n$ is lower than the number of predictors $p$.

2. Since each day the collected cats are selected randomly from the same population (of stray cats), each dataset comes from the same distribution of i.i.d. samples. Importantly, since the OLS assumptions are satisfied, the true model has additive uncorrelated Gaussian noise and therefore the $\hat{\beta}_w$ and $\hat{\beta}_s$ estimates also follow a (jointly) Gaussian distribution.

   (i) Case B.
   Uncorrelated predictors would lead to their estimates being uncorrelated Gaussians around their true value (because the OLS estimator is unbiased).

   (ii) Case C.
   In this case both predictors are almost similar. A higher value of $\hat{\beta}_w$ means a lower value of $\hat{\beta}_s$, simply because we need 'less' of $X_b$ when we use 'more' of $X_w$ and vice-versa. This leads to the two coefficients having a highly negative covariance.
   *See also: Lecture 3 - Linear Regression II, slide 30.*

   (iii) Case A.
   The reasoning is similar to the previous case: because the predictors are negatively correlated, lowering the contribution of one predictor leads to a higher contribution of the other. Therefore, the coefficient estimates will here have a highly positive covariance.

   - We leave out Case D.
     This coefficient distributions clearly *does not correspond* to a Gaussian, which means that the dataset on which these estimates were made cannot have satisfied the OLS assumptions.

3. (a) To verify whether a linear model is the right choice, we can draw and inspect the residuals plot: if the true relationship is non-linear it will reveal a U-like shape.
   *See also: Lecture 3 - Linear Regression II, slides 21-22.*

(b) One example is when an increase in advertisement budget for medium one affects (e.g. increases or decreases) the effectiveness of another.
For a positive effect, we can assume there is a synergy between the two media: having seen an ad in one would make seeing the ad in another more plausible or more persuasive.
For a negative effect, we can assume that the two media share some users, so advertising in one already covers a part of this population, rendering additionally showing it on another less effective.

(c) To test whether a predictor is contributing *significantly* to the outcome we can use a hypothesis test to reject the Null hypothesis that the corresponding coefficient is 0.
One such test is the $t$-test for the coefficient of $X^{99}_{\text{bf}}$.
*See also: Lecture 2 - Linear Regression, slide 15.*

4. (a) Let $\hat{\beta}'$ be the coefficient for the predictor that we get from a random dataset. Then, the statement in the box can be written as

$$\mathbb{P}\left(\hat{\beta}' > \hat{\beta} \middle| \beta = 0\right) < 2.2\% \,, \tag{1.1}$$

where $\beta$ is the true coefficient.
The name of this probability is the $p-$value of the hypothesis test based on the $z$-score to reject that the coefficient of $X$ is non-zero.

(b) The least squares assumptions hold and we are given the true variance of the estimate; therefore, we can compute the $z$-score of this estimate, which must follow a Gaussian distribution. This $z$-score is

$$z = \frac{\hat{\beta} - 0}{\text{SE}(\hat{\beta})} \,, \tag{1.2}$$

where $\text{SE}(\hat{\beta}) = \sqrt{\text{Var}(\hat{\beta})} = \sqrt{0.49} = 0.7$. The given probability is the integral of the right tail of the Gaussian probability density function for the part from $2\sigma$ and further to the right, as we can compute $2.1\% + 0.1\% = 2.2\%$. Therefore, we know that the $z$-score is equal to the value of 2.
We now plug these in equation (1.2) and we can now solve for $\hat{\beta}$, since

$$\frac{\hat{\beta}}{\text{SE}(\hat{\beta})} = 2 \iff \hat{\beta} = 2\,\text{SE}(\hat{\beta}) = 2 \cdot 0.7 = 1.4 \,.$$

*See also: Lecture 2 - Linear Regression, slide 15.*

## Problem 2 (Classification) **(10 points)**

1. A linear classifier that uses the predictors $X_1, X_2, X_1X_2, X_1^2$, and $X_2^2$ will have a decision boundary that falls into one out of 5 characteristic cases. Choose 4 out of the 5 cases, name them, and draw an example decision boundary for each. (1pt)

2. In Fig. 2 we show the decision boundaries for five different classification methods. Pair each of the boundaries to exactly one of the following classification methods and briefly explain each of your choices: (3pts)

   (i) LDA – Linear Discriminant Analysis
   (ii) QDA – Quadratic Discriminant Analysis
   (iii) LR – Logistic Regression
   (iv) 3-NN – $k$-Nearest neighbours with $k = 3$
   (v) SVC – Support Vector Classifier (hard margin, no kernel).

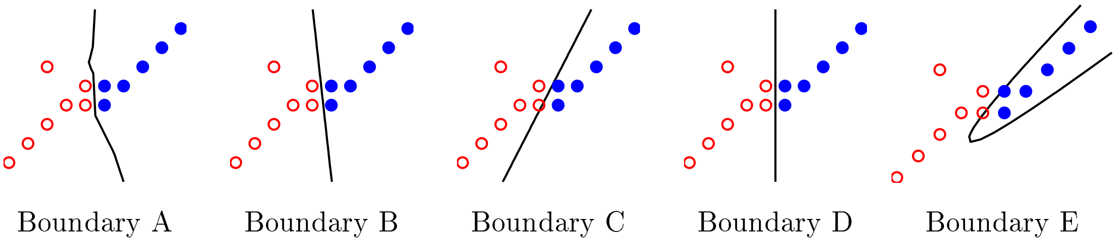


Figure 2: Different decision boundaries for the same dataset.

3. We consider the binary classification problem where based on a single predictor $X$ we want to classify samples into one out of two classes '+' and '−'. We know that

   - $P(X|Y = \text{'+'}) = \mathcal{N}(0, 1)$ is a Gaussian with zero mean and unit variance,
   - $P(X|Y = \text{'−'})$ is uniform over the interval $[-\alpha, \alpha]$ with some parameter $\alpha > 0$,
   - and $P(Y = \text{'+'}) = P(Y = \text{'−'})$ the classes are equally likely.

   Without performing extensive computation, answer the following questions.

   (a) Draw the decision boundary of the Bayes optimal classifier for $\alpha = 2.5$ on the real line of the graded axes in your answer booklet (page 5). You may use rounding to 1 decimal point. (1pt)
   (b) Briefly explain what happens to the decision boundary if $\alpha$ increases slightly. (1pt)
   (c) Briefly explain what happens to the decision boundary if the prior $P(Y = \text{'+'})$ increases slightly. (1pt)

   You may find useful the Gaussian probability density function shown in Fig. 3.

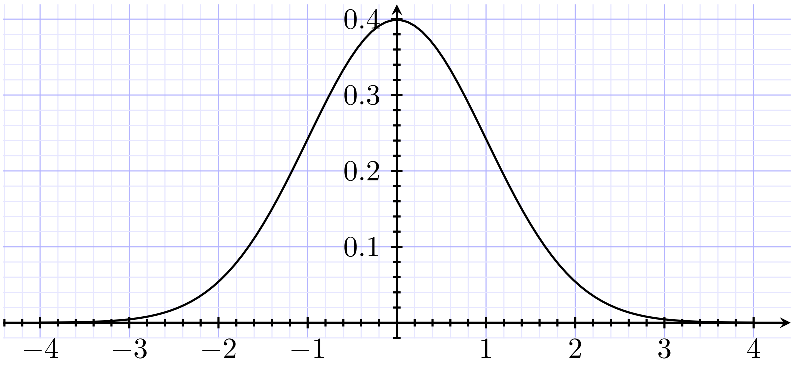


Figure 3: Probability density function of standard Gaussian $\mathcal{N}(0, 1)$.

4. You need to perform binary classification on the dataset shown in Fig. 4 that contains two predictors $X_1$ and $X_2$. You decide to use a support vector machine. For this, you consult for advice the not-yet-famous researcher Mara Saseche.

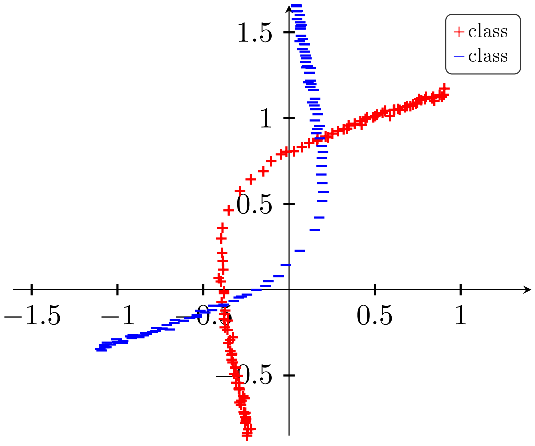
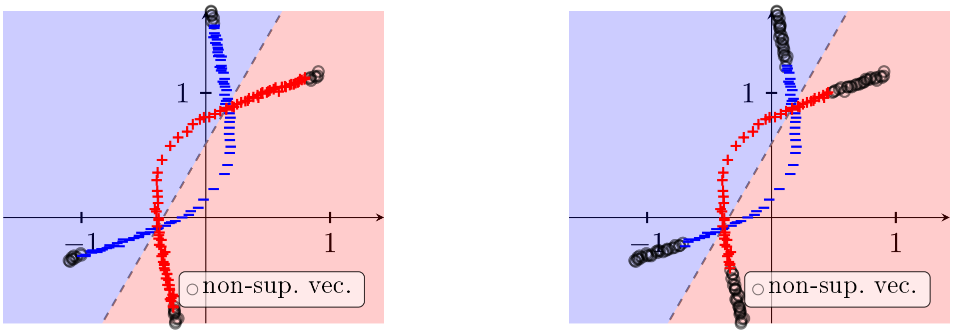


Figure 4: Dataset for support vector machine classification.

(a) You ask Mara to fit a hard margin classifier. Instead she fits the two soft margin classifiers shown in Fig. 5. Why couldn't Mara fulfill your request? How are the two models different? (1 pt)



(a) Model A

(b) Model B

Figure 5: Mara's two models. The dashed line shows the decision boundary. The background color shows the predicted class. The '+' and '−' denote data points that are support vectors. The black circles show data points that are *not* support vectors.

(b) You asked Mara for a better classifier. She came up with a support vector classifier for which the decision boundary is shown in Fig. 6. Briefly explain how she achieved this decision boundary. (1 pt)

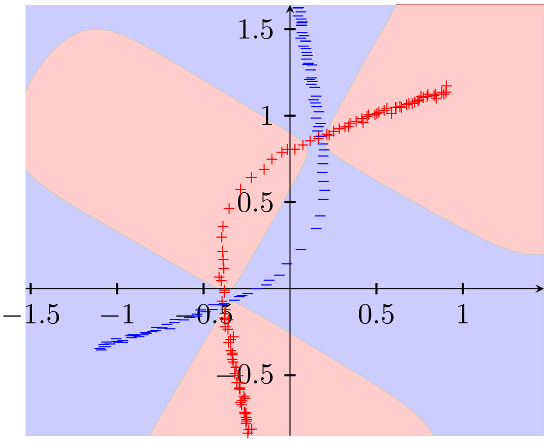


Figure 6: Support vector classifier with an improved decision boundary.

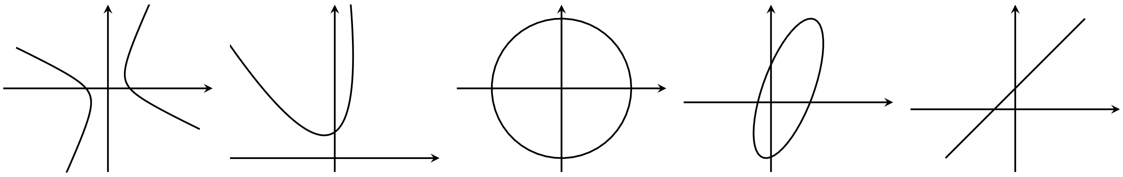
(c) As you discuss further, Mara informs you of the following correct statement. (1 pt)

> "*Using an SVM with a polynomial kernel of degree 2 is equivalent to using an SVM on a dataset that has been extended to additionally include predictors $X_1^2$, $X_1X_2$, and $X_2^2$.*"

Given the above, why would anyone still want to use a polynomial kernel?

*Solution.*

1. Using these terms allows us to create any quadratic boundary. These boundaries can be classified based on their shape as the parabola, hyberbolas, circles and ellipses, also including the linear boundary as a limit case.
   *Note: These shapes are also known as the conic sections.*



(a) Hyperbola (b) Parabola (c) Circle (d) Ellipse (e) Line

2. The rationale is to understand what each method does. The correct pairings are:

   (i) LDA - Boundary C
   LDA fits a Gaussian on each class, such that all share the same (co)variance, and then linearly separates their means. Here, the red class has an outlier on the upper left part, which largely shifts its mean upwards, thereby affecting the boundary a lot.

   (ii) QDA - Boundary E
   QDA fits a Gaussian on each class, such that all have their own (co)variance, and then linearly separates their means. This means QDA has a quadratic decision boundary. Here, due to the outlier, the variance of the red class is much higher, which drives the method to "squeeze" the decision boundary of the blue class to the more compact quadratically separated part.

   (iii) LR - Boundary B
   Logistic regression takes all points into consideration, but is not as prone to outliers as e.g. LDA, as it weights points further from the decision boundary less weighted. Here, it correctly detects the trend of the points near the boundary.

   (iv) 3-NN - Boundary A
   k-NN is a non parametric and therefore quite flexible method, and is the only one here that can achieve this kind of non-linear and non-quadratic decision boundary.

   (v) SVC - Boundary D
   Since the classes are linearly separable, the hard-margin classifier only looks at the support vectors. Here, these are just the four central points, which together define the perpendicular decision boundary.

3. This is a simplification of *Problem 2 of Assignment Sheet 2.*

   (a) Since the class priors are equal, we only need to compare the likelihoods.
   Although not necessary, this can be shown as below:

$$\mathbb{P}\left(Y = \text{`}+\text{'}|X\right) > \mathbb{P}\left(Y = \text{`}-\text{'}|X\right) \qquad \Longleftrightarrow \quad \text{using Bayes' rule}$$

$$\frac{\mathbb{P}(X|Y=\text{`}+\text{'})\,\mathbb{P}(Y=\text{`}+\text{'})}{\mathbb{P}(X)} > \frac{\mathbb{P}(X|Y=\text{`}-\text{'})\,\cancel{\mathbb{P}(Y=\text{`}-\text{'})}}{\mathbb{P}(X)} \iff \text{equal class priors}$$

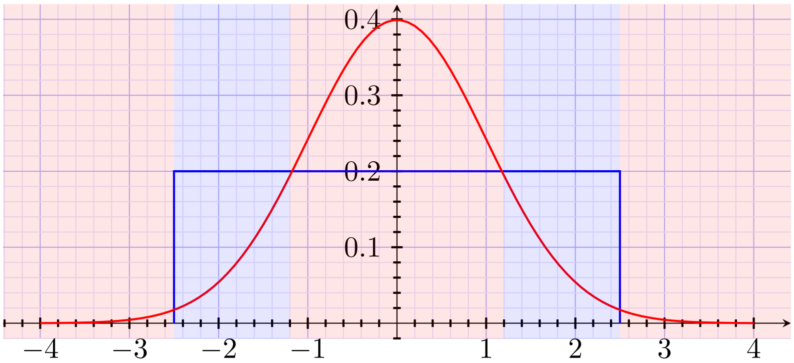
$$\frac{\mathbb{P}(X|Y=\text{`}+\text{'})\,\cancel{\mathbb{P}(Y=\text{`}+\text{'})}}{\mathbb{P}(X)} > \frac{\mathbb{P}(X|Y=\text{`}-\text{'})\,\cancel{\mathbb{P}(Y=\text{`}-\text{'})}}{\mathbb{P}(X)} \iff \text{multiply by } \mathbb{P}(X) > 0$$

$$\mathbb{P}(X|Y=\text{`}+\text{'}) > \mathbb{P}(X|Y=\text{`}-\text{'})\,.$$

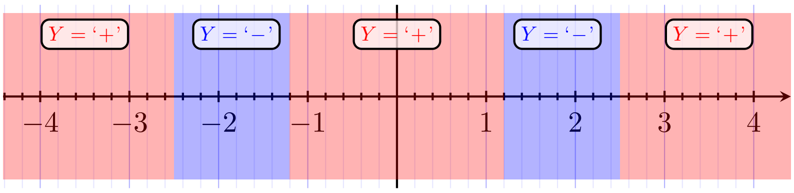
The density of the uniform distribution is some constant $c$ within $[-2.5, 2.5]$. To find $c$ we must compute the value that makes the area of the defined rectangle equal to 1

$$c \cdot (2.5 - (-2.5)) = 1 \iff c = \frac{1}{5} = 0.2\,.$$

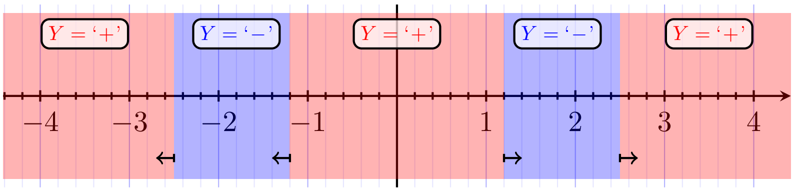
We now plot both densities one on top of the other and select the most likely class.



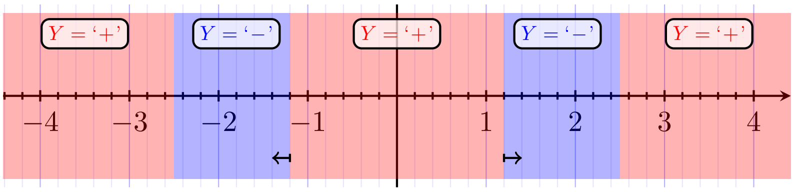
Therefore, the resulting boundary is shown below.



(b) If $\alpha$ increases slightly, the density $c$ will become lower, and therefore, the central region will expand.
At the same time, the support of the negative class will increase, so the outer regions $Y = \text{`}+\text{'}$ will move further from the origin to reflect the increased $\alpha$.



(c) If the prior $\mathbb{P}(Y = \text{`}+\text{'})$ increases slightly, the density of the Gaussian will be scaled up during the comparison. Although a small change will not affect the outermost region, it will lead to the expansion of the innermost one.



4. (a) A hard margin classifier requires the two classes to be linearly separable. As this is not the case here, Mara could not provide this classifier.
Instead, she could compute soft-margin classifiers with different budget parameters $C$, as this does not require strict linear separability. Model A has a higher budget $C$ than that of Model B, which we can see from the fact that it contains many more support vectors than the latter.

(b) To achieve this boundary she used a (non-linear) kernel.
The kernel she used was the (Gaussian) radial basis, which allows the complex boundary we see here. In contrast, the polynomial kernel in the 2 dimensions that we have here would just give a quadratic boundary, which is much simpler than the one we see here.

(c) Even though the two approaches are the same in theory, in practice it would require too many resources to extend the dataset with more features and for higher degrees. More specifically, with $n$ features and a degree of $d$ we would need to extend our dataset to a total of $n^d$ predictors.
Instead, the polynomial kernel allows for a much more efficient computation; for this we only compute an inner product of the two feature vectors raised to the power of $d$. This is a good example where we do not need to create the entire feature space for the computation of the kernel.
*See also Lecture 12 slides 23-24.*

## PROBLEM 3 (TREES AND FORESTS) **(10 points)**

1. Consider the following dataset of two predictors $X_1, X_2$ and one target $Y$. (1pt)

| $X_1$ | $X_2$ | $Y$ |
|---|---|---|
| 1 | 1 | 1 |
| 1 | 2 | 1.5 |
| 2 | 1 | 11 |
| 2 | 2 | 10.5 |

Construct a regression tree for the data such that each leaf contains precisely one data point. Use the approximation $(a-x)^2 \approx a^2 - 2ax$ for $x < 1$ when computing the MSE gains.

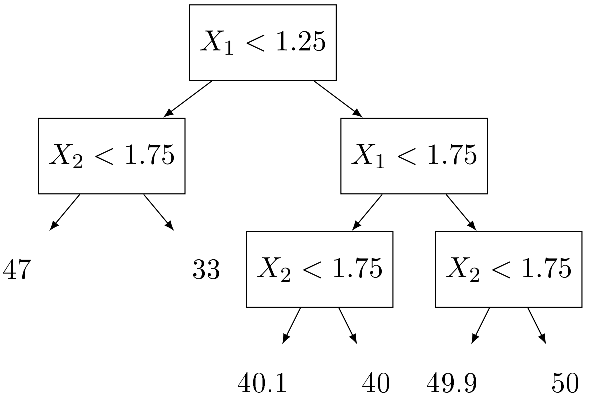
2. Consider the regression tree shown in Fig. 8a.

(a) Draw the corresponding regions and indicate the value in each region. (1pt)

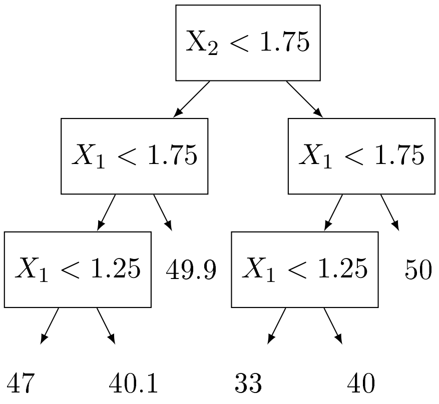
(b) Consider the right-most split on $X_2 < 1.75$ in the tree. In terms of generalization, under which conditions does it makes sense to have this node in the tree? What would lead to having this node in the tree even when it does not make sense? (1pt)

Now consider both regression trees shown in Fig. 8.

(c) For both pruning and constraining tree depth, explain whether it would work better, worse, or equally well for the left tree vs. the right tree. Based on this, explain the shortcomings of both these approaches for constraining flexibility in regression trees. Explain your reasoning. (3pts)



(a) Regression tree 1



(b) Regression tree 2

Figure 8: Two equivalent regression trees.

3. World-famous ensemble learning researchers Roav Schreund and Yobert Fapire say that modeling data with only a single regression tree is a bad idea, and that we should instead use an ensemble of trees.

   (a) Explain how Bagging works, how Random Forests work, and how these two differ from each other. (1pt)

   (b) Explain how Boosting works, and how it differs from Bagging. (1pt)

   (c) Explain how variable importance is computed for a Random Forest. Can we use the same approach for Boosted Trees? Why (not)? (1pt)

4. Not-yet-famous researcher Kavid Daltenpoth makes the following statement: (1pt)

   *"Bagging, Boosting, and Random Forests are all linear models".*

   Give one argument or example in favor, and one example or argument against this statement.

*Solution.*

1. (a) Looking at the data, we note that the mean across all data points $Y$ is $\frac{1+1.5+11+10.5}{4} = 6$. The approximate MSE using the formula given in the exercise is therefore $\frac{5^2+4.5^2}{2} = \frac{25}{2} + \frac{5^2-2\cdot5\cdot0.5}{2} = 12.5 - 2.5 = 10$.
Considering the two splits $X_1 < 1.5$ and $X_2 < 1.5$, we see that the latter produces the two groups $\{1, 11\}$ and $\{1.5, 10.5\}$, both of which have a mean of 6 again. The split on $X_2$ therefore has *zero* gain. In comparison, the split on $X_1 < 1.5$ produces the groups $\{1, 1.5\}$ and $\{10.5, 11\}$ with means 1.25, respectively 10.75. The MSE for this split is therefore $0.25^2$ for a gain of $10 - 0.25^2$, which is clearly much larger than 0. The second split on each branch is trivial, and the resulting tree is given in Fig. 9

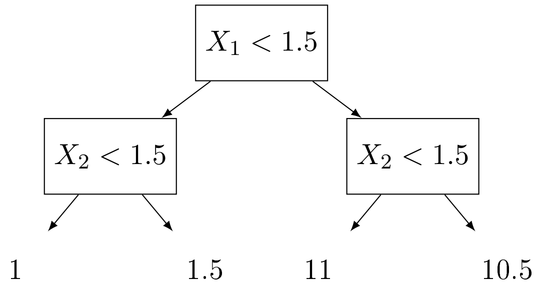


Figure 9: Regression Tree.

2. (a) The resulting regions (restricted to $[0, 2]^2$) are given by Fig. 10

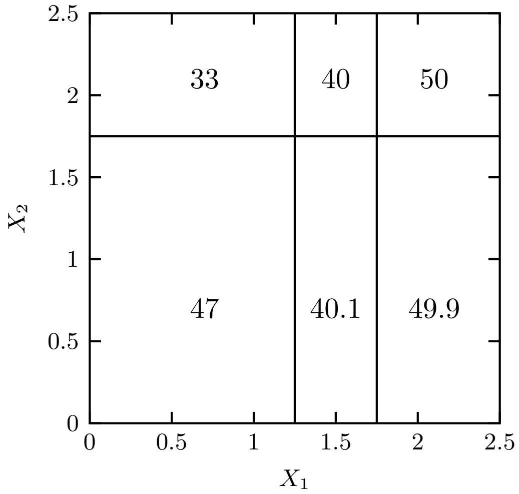


Figure 10: Resulting decision regions.

(b) The split between values 49.9 and 50 is meaningful if there is either 1) enough of a difference between the two values to justify the distinction (think of elections), or 2) enough data to make such a split significant. When neither of these two conditions apply, however, such a split is likely due to overfitting.

(c) Both pruning and depth restriction would work better for Regression tree 1 shown in Fig. 8a. Since the tree in Fig. 8b has very different values in its leaves on the same level, trying to prune this tree would result in a large increase in MSE and would therefore likely not work. Tree 1, however, has very similar leaf values, pruning therefore likely to work well.
Similarly, restricting the depth would also work better for the Regression tree 1. In both trees, once we decide on the top-level split, the remaining splits maximizing MSE gains are easily determined, and therefore restricting the depth to 2 would lead to the same result as pruning the leaves so that the same reasons apply.

3. (a) Bagging works by taking $B$ independent bootstrap samples from the original dataset and fitting a tree $T_b$ to each of these datasets, and predicting the value for a given value $x_0$ by averaging the trees: $\hat{y}_0 = \sum_{b=1}^{B} T_b(x_0)$. Since bootstrap samples have large overlap, these trees are generally highly correlated, however. Random Forests therefore in addition use only a random subset of $p$ predictors for each of the trees to reduce this correlation.

(b) Bagging works by averaging trees trained independently from one another on different datasets. In contrast, boosting fits a sequence of models. It iteratively fits a new model to the data, and then changing the weights of each data point according to its residual, i.e., how (not) well the model predicts that data point. The newly fit model therefore tries to predict those points which are not predicted well by the previous model.

(c) The importance of a predictor for a single tree is computed as

$$I_l^2(T) = \sum_{t=1}^{J-1} g_t^2 I(v(t) = l)\,,$$

where $g_t$ is the gain from the split at node $t$ in tree $T$ and $v(t)$ the predictor we split on. That is, we simply measure the total improvement in gain due to splits on predictor $l$. When we have multiple trees $T$ in a random forest, we can simply average the gains for each of its individual trees.
For boosting, we can take precisely the same approach of averaging the gains due to splitting on the predictor $l$.

4. **In favor**: Since bagging, boosting and random forests are simply averages over multiple trees, it suffices to give an argument that trees can be considered as linear models. Since every tree basically amounts to assigning different (constant) $c_i$ values to different regions $R_i$ defined by the tree $T$, we can write the prediction at a point $x_0$ as

$$T(x_0) = \sum_i c_i I(x_0 \in R_i)\,.$$

A regression tree can therefore be considered linear regression with basis functions $I(\cdot \in R)$ for all possible regions $R$ that can be constructed by regression trees.

**Against**: The set of all possible regions is uncountably infinite and therefore makes it impossible to actually do such a regression. The task of determining *which* regions need to be considered is precisely part of the problems that regression trees solve.

## PROBLEM 4 (MODEL SELECTION) **(10 points)**

1. We are given two datasets, one of $n = 10$, and a second of $n = 1\,000$ samples, over ten predictors $X_1, \dots, X_{10}$ and one continuous-valued target variable $Y$. We want to find out which are the relevant predictors for $Y$.

   (a) We fit a linear regression model using each possible subset of predictors. In Fig. 11, we show the BIC score of the regression model using the $k$ best predictors. Which line corresponds to the small and which to the large dataset? Explain. (1pt)

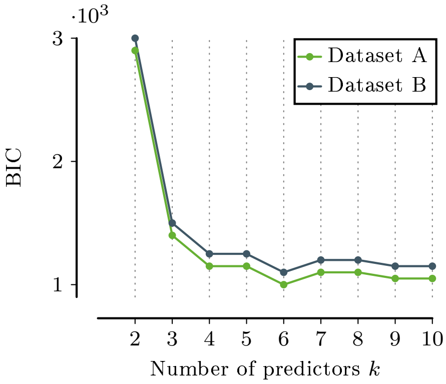


Figure 11: BIC scores of linear regression models that use the $k$ best predictors.

   (b) Based on Fig. 11, how many predictors would you choose and why? (1pt)

   (c) Explain how BIC differs from AIC. Based on this, does the choice between BIC and AIC matter more for the small or for the large dataset? (1pt)

   (d) Briefly explain how we can use each of the following methods to select relevant predictors for $Y$, and give one advantage and one disadvantage: (3pts)

   - subset selection,
   - cross validation,
   - shrinkage.

2. Consider the following formulation to find the linear regression parameter $\hat{\beta} \in \mathbb{R}^d$,

$$\hat{\beta} = \arg\min_\beta \sum_{i=1}^{n} (y_i - \beta x_i)^2 \qquad \text{subject to} \quad \|\beta\|_2^2 \le K\,.$$

   (a) Explain the effect of the constraint in the above objective. (1pt)

   (b) How does the above formulation differ from OLS in terms of the number of parameters and in terms of the degrees of freedom? (1pt)

   (c) Consider varying the value of parameter $K$. What would be the effect on the bias and variance of the model? (1pt)

(d) How could we modify the constraint to perform subset selection? (1pt)

*Solution.*

1. (a) BIC is given by

$$\mathrm{BIC} = K \log n - 2 \log L \ .$$

In the case of linear models, the number of parameters $K$ corresponds simply to the number of predictors $k$ and does not differ between the small and large dataset. Due to the logarithmic term in $n$, BIC may be slightly larger for large sample sizes, unless there is a large enough difference in the likelihood $L$. In our example, we might thus decide that dataset A is the smaller one and dataset B the larger one. *Note:* for more complex models than a linear one, the number of parameters $k$ of the best models will also depend on the dataset. For example, for very small sample size $n = 10$, the models we infer from the dataset may be misspecified and contain e.g. too many parameters.

(b) We choose the model with the minimal BIC value, which here has 6 predictors. Mind the difference to the cross-validation rule, where we choose the simplest model within one standard deviation of the minimal one, or the elbow rule.

(c) AIC is given by

$$\mathrm{AIC} = 2k - 2 \log L \ .$$

BIC and AIC differ in the penalty they put on the model parameters. In AIC it is independent of the sample size, and the number of datapoints does not directly enter the AIC score. Comparing two datasets of different sizes, the AIC scores only differ by the likelihood term and may be more similar to each other than the BIC scores. *Note:* Again, we need to keep in mind that one dataset is much too small ($n = 10$) to expect reasonable model parameters. Hence, for nonlinear models with varying number of parameters $k$ we may get too many (or too few) parameters. Then, the data size can indirectly influence AIC.

(d) The three methods are related to feature selection as follows.

- **Subset Selection**: We consider each possible subset of predictors and choose the one that is best under a given metric (such as the CV error). While this gives us an exact result and does not require the specification of any hyperparameters, the approach is limited to small numbers $p$ of predictors due to its heavy computational cost in $\mathcal{O}(2^p)$. *Note:* Distinct from stepwise selection which is a greedy approach.
- **Cross Validation**: This is one option for choosing the optimal set of predictors, i.e. can be used to compare the models found with best subset or stepwise selection. We perform cross validation and choose the simplest model within one standard error of the best one. As an advantage, cross validation takes different splits of the data into account to give a more robust estimate of model performance. Besides feature selection, it is also applicable to other use cases, e.g. selecting tuning parameters in regularization. We need to set the hyperparameter $k$; it adds a computational overhead, especially for large $k$; and it may not work well with imbalanced datasets.

- **Shrinkage Methods**: We can use Lasso regularization to drive the coefficients of irrelevant features to zero. As an advantage, using Lasso we can directly remove the effect of irrelevant features on the model prediction, without the need for postprocessing. A disadvantage is that you need to choose a tuning parameter, that they are not necessarily applicable to non-linear models, and may be sensitive to collinearity.

2. The formulation is

$$\hat{\beta} = \arg\min_\beta \sum_{i=1}^{n} (y_i - \beta x_i)^2 \text{ subject to } \|\beta\|_2^2 \leq K\,.$$

(a) This is simply a constrained formulation of ridge regression, which we have seen in its Lagrange formulation in the lecture.

(b) Since the parameter $K$ is a hyperparameter in the constraint and not part of the model fitting itself, the number of parameters is the same as in OLS, $\beta$. However, we do have a reduction in the degrees of freedom compared to OLS:

(c) Choosing a smaller $K$ increases the regularization strength, which introduces some bias as it restricts the degrees of freedom, but has potentially less risk of overfitting and hence lower variance. Similarly, increasing $K$ moves the objective closer to OLS, lowering the bias and increasing the variance.

(d) To turn this into subset selection, we can replace $\|\beta\|_2^2$ by an indicator function as follows,

$$\hat{\beta} = \arg\min_\beta \sum_{i=1}^{n} (y_i - \beta x_i)^2 \text{ subject to } I(\beta_j \neq 0) \leq K\,,$$

where $I(\beta_j \neq 0)$ is 1 when $\beta_j \neq 0$ and 0 otherwise. For each predictor $X_j$, we include it in the model whenever the corresponding coefficient $\beta_j$ is nonzero. Hence, the above constraint amounts to using a subset of $K$ predictors. *Note*: Replacing the L2 norm by L0 instead would simply turn this into Lasso regression. Also, we can state the above also as a Lagrange formulation with the same indicator function and a penalty $\lambda$ for including more predictors.

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

1. Not-yet-famous researcher Kanis Jalofolias is still interested in cats. Looking at the plot in Fig. 12, he is starting to suspect that not all cats are created equal and wants to determine which cats are similar to each other.

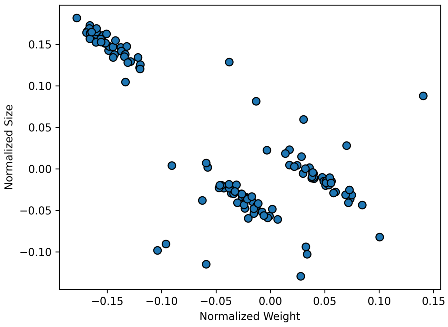


Figure 12: A scatter plot of two different predictors for different cats.

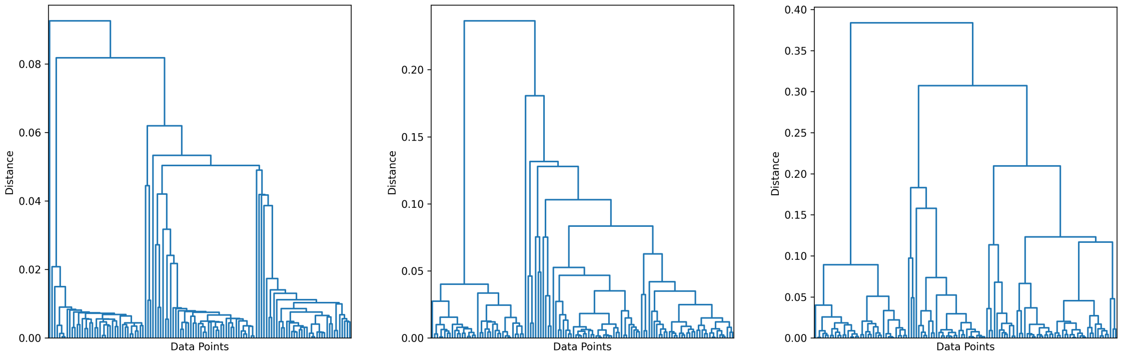
   (a) Explain how the $k$-means clustering algorithm works. Would you expect it to work well on the data depicted in Fig. 12? Why (not)? (1pt)

   (b) Kavid and Kanis both run $k$-means on the same data, with the same distance measure, and the same value for $k$. They get different results. What happened? (1pt)

2. Catvaid suggests that instead of using $k$-means, Kavid and Kanis should use hierarchical clustering.

   (a) Explain how hierarchical clustering differs from $k$-means clustering. (1pt)

   (b) For the three dendrograms in Fig. 13, explain which one corresponds to single, which one to average, and which one to complete linkage. (2pts)



(a) Dendrogram A (b) Dendrogram B (c) Dendrogram C

Figure 13: Dendrogram plots for different linkage methods.

3. Unhappy with the progress in his quest to understand cats, Kanis measures the genotype of every cat he comes across, resulting in a dataset of $d = 20\,000$ features (genes) and $n = 10\,000$ samples (cats).

   (a) Having gathered all this data, Kanis starts scratching his head, unsure how to deal with it. Explain the problem with the data set he has collected. (1pt)

   (b) In a discussion with Kavid, the two decide to use PCA to reduce the number of features. Kavid suggests they should use PCA *before* clustering the data, while Kanis is insistent on clustering the data first and then using PCA to visualize the resulting clusters. Give arguments both in favor and against either approach. (2pts)

   (c) Mara tells Kanis and Kavid that they are both wrong and should use $t$-SNE instead. Succinctly explain how $t$-SNE works and how it differs from PCA. (2pts)

*Solution.*

1. (a) $k$-means clustering tries to find $k$ clusters around $k$ cluster means such that the sum of squared distances from each data point to its closest cluster mean is minimized. It would not work well on this data-set because the minimization of squared distances implicitly assumes that the data is distributed according to a mixture spherical Gaussian distributions. However, the clusters in this data are elongated and contain multiple outliers, violating these assumptions.

   (b) Since $k$-means clustering starts with a random initialization of the cluster means, two runs starting with different initializations can result in very different results.

2. (a) $k$-means clustering is a parametric method tries to find one (local) optimum for clustering the data in a fixed number of $k$ different clusters. In contrast, hierarchical clustering is a non-parametric bottom-up method starting with clusters on each individual data points and iteratively merges those clusters which have the lowest distance between them, as defined by the linkage criterion used. In particular, it returns (locally) optimal clusters for each value of $k = 1, \ldots, n$.

   (b) The easiest way to distinguish between the linkage methods based on the dendrograms for the same data set is to look at the distances. In general, for any two given clusters we have that the distances measured compare as follows: complete linkage $\geq$ average linkage $\geq$ single linkage. As such, the order from single to complete linkage is precisely left to right.

   A different way of seeing this is to look how quickly each of the plots creates larger clusters. The leftmost plot creates larger clusters relatively quickly, whereas the rightmost plot mostly starts out by merging individual data points, with the middle plot intermediate between the two. This corresponds precisely to the different linkage methods we argued for above: since single linkage takes into account only the *smallest* distance between pairs of points in clusters, it is very easy for it to form large clusters. In contrast, for complete linkage, taking into account the *largest* distance between pairs of points, it is much more difficult to form larger clusters. Average linkage is intermediate between these two and therefore results in the intermediate dendrogram.

3. (a) The problem here is that we have only 10 000 data points for 20 000 features, so that our data is very sparsely distributed and the curse of dimensionality is in full effect.

   (b) Using PCA before clustering: If the assumptions underlying PCA are fulfilled then using it before clustering may be a good idea since most of the relevant information would still be contained in the lower dimensional data set. In contrast, if the assumptions underlying PCA are not fulfilled, then no matter how good our clustering algorithm may be, we cannot expect any good results to come of using it on the output of PCA.

   Using PCA after clustering: As noted above, if the assumptions behind PCA are faulty then no matter the clustering algorithm we will not get good results. By clustering first and using PCA after, we avoid this potential problem.

However, if the PCA assumptions are satisfied then clustering first and using PCA afterwards then we are obtaining potentially suboptimal clusters by clustering in an extremely high-dimensional space when clustering in the much lower-dimensional space would have sufficed.

(c) $t$-SNE takes a matrix of pair-wise distances $d_{ij}$ between points in high-dimensional space and tries to find an embedding in a lower-dimensional space with pairwise distances $q_{ij}$ which minimizes the difference between the distances $d_{ij}$ and $q_{ij}$ by minimizing the KL divergence between the two.

The main differences are that PCA is linear and $t$-SNE is not, and furthermore PCA fits one global model to the data while $t$-SNE tries to maintain local distances.