5 Model Selection

Exercise 5.1

Consider the following snippet of code taking as input an `N` $\times$ `M` matrix `X` of data points and an `N`-dimensional vector `y` of binary targets, where `N` is the number samples, `M` is the number of features. Are the subsequent statements true or false? Provide motivations for your answers.

```
1  import numpy as np
2  classifier = LogisiticRegression()
3  classifier.fit(X, y)
4  yhat = classifier.predict(X)
5  accuracies = [sum(yhat = = y) / N]
6  for i in range(M):
7     Xi = np.delete(X, i, axis=1)
8     classifier.fit(Xi, y)
9     yhat = classifier.predict(Xi)
10    accuracies.append(sum(yhat = = y) / N)
```

1. This snippet of code implements a portion of a well-known model selection procedure.
2. After having run the reported snippet, one should keep the classifier that led to the maximum value in the list of accuracies.
3. The classifier trained at line 3 is likely to suffer a larger bias than the classifiers trained at line 8.
4. The time of computation we need to run the snippet of code scales linearly with `M`.

Exercise 5.2

Answer to the following questions regarding feature selection. Provide motivation for your answers.

1. You have been asked to implement a feature selection process on a system with very limited computational resources. Would you opt for a filter approach or for a wrapper approach?

2. You have been asked to implement a feature selection process to improve as much as possible the performance of your model. Would you opt for a filter approach or for a wrapper approach?

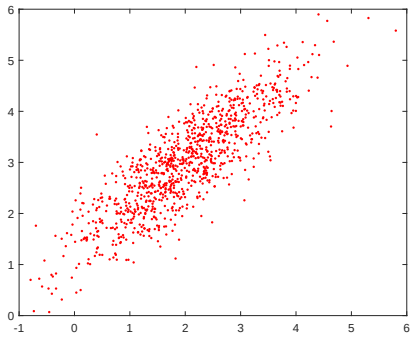
3. If you want to rank (in order of importance) the features of a classification problem, which kind of feature selection process would you use among the ones presented in the course?

4. You trained two models on a problem with 5 features: Model A using all the 5 features and Model B using only 3 features. Assuming they have similar performances on the training set, do you expect Model A to perform better on the test set?

5. You trained two models on a problem with 9 features: Model A using only 5 features and Model B using all the 9 features. Do you expect Model A to have a smaller training error than Model B?

Exercise 5.3

Consider the following dataset:



Draw the direction of the principal components and provide an approximate and consistent guess of the values of the loadings. Are they unique?

Exercise 5.4

Consider the following statement regarding PCA and tell if they are true or false. Provide motivation for your answers.

1. Even if all the input features are on very similar scales, we should still perform mean normalization (so that each feature has zero mean) before running PCA.

2. Given only scores t_i and the loadings W, there is no way to reconstruct any rea-

sonable approximation to x_i.

3. Given input data $x_i \in \mathbb{R}^d$, it makes sense to run PCA only with values of k that satisfy $k \leq d$.

4. PCA is susceptible to local optima, thus trying multiple random initializations may help.

Exercise 5.5

You are analysing a dataset where each observation is an age, height, length, and width of a turtle. You want to know if the data can be well described by fewer than four dimensions, so you decide to do Principal Component Analysis. Which of the following is most likely to be the loadings of the first Principal Component?

1. $(1,\ 1,\ 1,\ 1)$
2. $(0.5,\ 0.5,\ 0.5,\ 0.5)$
3. $(0.71,\ -0.71,\ 0,\ 0)$
4. $(1,\ -1,\ -1,\ -1)$

Provide motivations for your answer.

Exercise 5.6

Consider the following snippet of code taking as input a dataset `X` having 100 samples with 5 features. Are the subsequent statements true or false? Motivate your answers.

```
1 import numpy as np
2 X_tilde = X - np.mean(X, axis=0)
3 C = np.dot(X_tilde.T, X_tilde)
4 eigenvalues, W = np.linalg.eig(C)
5 T = np.dot(X_tilde, W[:,:2])
```

1. The code snippet above is implementing a feature selection technique.
2. Line 2 is unnecessary if the data in `X` are scaled.
3. A model trained with the inputs `T` is likely to display a lower bias than a model trained with inputs `X`.
4. It is possible to recover, by computing $X = W[:, :2]^\top \cdot T$, the original dataset `X` from `T`.

Exercise 5.7

A KNN classifier classifies a new data point by applying a majority voting among its K-Nearest Neighbour. These are the available data points in the dataset:

$$x_0 = (-2, -3, 0), y_0 = 1 \qquad x_1 = (-2, 2, 1), y_1 = 1$$
$$x_2 = (-2, -1, -3), y_2 = 1 \qquad x_3 = (2, 1, -4), y_3 = 1$$
$$x_4 = (2, -3, 2), y_4 = 1 \qquad x_5 = (1, 2, 2), y_5 = 0$$
$$x_6 = (-1, 1, -2), y_6 = 0 \qquad x_7 = (-1, 2, 2), y_7 = 1$$
$$x_8 = (1, -2, 0), y_8 = 0 \qquad x_9 = (-3, -3, -2), y_9 = 1$$

1. Classify the new point $x = (0, 0, -1)$ according to a KNN classifier trained on the dataset reported below, assuming $K = 3$ and using the Euclidean distance;
2. What happens if we use $K = 10$ instead? Do you think it is a good idea to choose such a parameter (hint: two pros if it is a good idea or two cons if it is not);
3. Suggest a technique to set the parameter K.

Exercise 5.8

State whether the following claims about Bagging and Boosting are true or false, motivating your answers:

1. Since Boosting and Bagging are ensemble methods, they can be both parallelized.
2. Bagging should be applied with weak learners.
3. The central idea of Boosting consists in using bootstrapping.
4. It is not a good idea to use Boosting with a deep neural network as a base learner.

Exercise 5.9

Which of the following is a reasonable way to select the number of principal components k in a datset with N samples?

1. Choose k to be 99% of N, i.e., $k = \lceil 0.99N \rceil$;
2. Choose the value of k that minimizes the approximation error $\sum_{i=1}^{N} ||x_i - \hat{x}_i||_2^2$;
3. Choose k to be the smallest value so that at least 99% of the variance is retained;
4. Choose k to be the smallest value so that at least 1% of the variance is retained;
5. Identify the elbow of the cumulated variance function.

What changes if the purpose of PCA is visualization?