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?

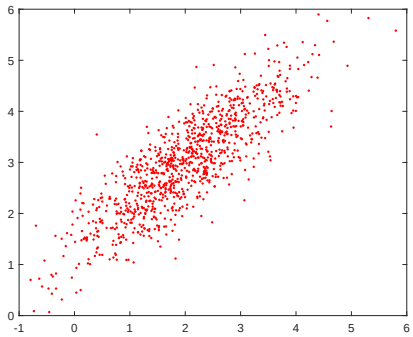
3. If you want to rank (in order of importance) the features of a classification problem, which kind of feature selection process would you use among the ones presented in the course?

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?

Answers

Answer of exercise 5.1

1. TRUE the code implements the first iteration of a typical *backward feature selection* procedure, which is used for model selection. However, some aspects are missing in the snippet, such as the computation of the accuracy on a validation set or cross-validation.

2. FALSE model selection cannot be performed on the training accuracy, otherwise we would always select the model that keeps all the `M` features.

3. FALSE the classifier trained on the full set of features is likely to suffer a larger variance and a lower bias w.r.t. the classifiers trained on a subset of the features.

4. FALSE whereas we need to run the loop 6-10 exactly `M` times, the time of computation we need to run the fit function of the classifier grows with `M` as well, being the logistic regression a parametric method.

Answer of exercise 5.2

1. FILTER because a wrapper approach involves solving an optimization problem that requires training several models. In contrast, filter approaches only require to compute statistics on the features.

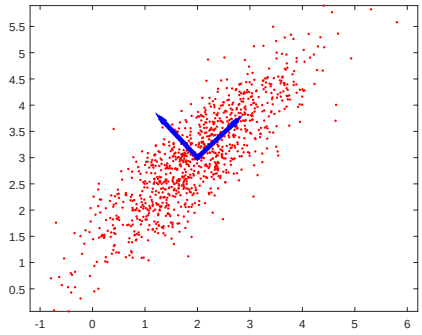
2. WRAPPER because filter approaches assume features are independent and might not find the best subset.

3. FILTER because it involves computing a metric for each feature (e.g., correlation or information gain) such that on the basis of this metric features can be ranked and selected from the most relevant to the least one.

4. NO Model A is more complex and hence has a larger variance and probability of overfitting training data. So it would probably result in a worse test error.

5. NO Model A is simpler and will probably have a larger bias resulting in a larger training error.

Answer of exercise 5.3



The computed principal components loadings are:

$$\begin{matrix} 0.7287 & -0.6849 \\ 0.6849 & 0.7287 \end{matrix}$$

and a reasonable guess would be:

$$\begin{matrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{matrix}$$

The properties satisfied by the principal components are the unity norm and that they are orthogonal.

Answer of exercise 5.4

1. TRUE Since the principal components are identifying the directions where the most of the variance of the data is present, where the directions is defined as a vector with tail in the origin, we should remove the mean values for each component in order to identify correctly these directions.

2. FALSE By applying again the loadings matrix W to the scores t_i, thanks to the orthogonality property of W, we are able to reconstruct perfectly the original mean normalized vectors. If we want to reconstruct the original vectors we should also store the mean values for each dimension.

3. TRUE Running it with $k = d$ is possible but usually not helpful and $k > d$ does not make sense.

4. FALSE There is no source of randomization and no initialization point in the algorithm to perform PCA.

Answer of exercise 5.5

Options 1 and 4 cannot be right because the sum of the squared loadings (i.e., their norm) exceeds 1. Options 2 and 3 are both reasonable options.

The second option is most likely correct because we expect this particular set of four variables to be positively correlated with each other in the first principal component. Note that it is fairly common for the loadings of the first principal component to all have the same sign. In such a case, the principal component is often referred to as a size component.

Answer of exercise 5.6

1. FALSE the code snippet is implementing the PCA, which is a feature extraction technique.
2. FALSE the data in `X` has to be centered anyway to compute the covariance matrix.
3. FALSE the dataset `T` has lower dimensionality than the dataset `X`. Thus, a model trained on `T` is likely to have a larger bias but a smaller variance.
4. FALSE in this way we are minimizing the reconstruction error, but it would be greater than zero in general.

Answer of exercise 5.7

1. Let us compute the Euclidean distances

$$
\begin{aligned}
d(x, x_0) &= d((0,0,-1),(-2,-3,0)) = \sqrt{(4+9+1)} = \sqrt{(14)} = 3.74 \\
d(x, x_1) &= d((0,0,-1),(-2,2,1)) = \sqrt{(4+4+4)} = \sqrt{(12)} = 3.46 \\
d(x, x_2) &= d((0,0,-1),(-2,-1,-3)) = \sqrt{(4+1+4)} = \sqrt{(9)} = 3.0 \\
d(x, x_3) &= d((0,0,-1),(2,1,-4)) = \sqrt{(4+1+9)} = \sqrt{(14)} = 3.74 \\
d(x, x_4) &= d((0,0,-1),(2,-3,2)) = \sqrt{(4+9+9)} = \sqrt{(22)} = 4.69 \\
d(x, x_5) &= d((0,0,-1),(1,2,2)) = \sqrt{(1+4+9)} = \sqrt{(14)} = 3.74 \\
d(x, x_6) &= d((0,0,-1),(-1,1,-2)) = \sqrt{(1+1+1)} = \sqrt{(3)} = 1.73 \\
d(x, x_7) &= d((0,0,-1),(-1,2,2)) = \sqrt{(1+4+9)} = \sqrt{(14)} = 3.74 \\
d(x, x_8) &= d((0,0,-1),(1,-2,0)) = \sqrt{(1+4+1)} = \sqrt{(6)} = 2.45 \\
d(x, x_9) &= d((0,0,-1),(-3,-3,-2)) = \sqrt{(9+9+1)} = \sqrt{(19)} = 4.36
\end{aligned}
$$

   The closest points are (x_6, x_8, x_2), thus we perform majority voting within $(y_6, y_8, y_2) = (0, 0, 1)$, which classifies the new point as 0.

2. It is not a good idea since in this case we would have exactly the same results for each new datapoint. Moreover, in general, it is a good practice to select an odd value for the K parameter to avoid ties when doing majority voting.

3. We can use cross-validation (k-fold) to select the K having the lowest expected error.

Answer of exercise 5.8

1. FALSE only Bagging can be parallelized, since training is done on different datasets, while Boosting is sequential by nature.

2. FALSE weak learners are good candidate for Boosting, since they have low variance. Typically one uses instead bagging when more complex and unstable learners are needed, to reduce their variance.

3. FALSE bootstrapping is mainly used in Bagging, whose name derived indeed from "bootstrap aggregation".

4. TRUE it is not a good idea to do that, since deep neural networks are very complex predictor, which can have large variance. Therefore, you may not succeed in lowering bias without increasing variance. Moreover, since you need to train the network multiple times, the procedure may require a lot of time.

Answer of exercise 5.9

1. FALSE This way we could either include principal components which provides explanation for small amount of variance or exclude some of the most important ones.

2. FALSE This would mean to include all the principal components, which are able to perfectly reconstruct the original dataset;

3. TRUE The cumulated variance of the principal components provides us an estimates on how much of the variance of the original dataset is considered. Keeping an high percentage of it would imply that we are discarding components which does not vary too much or noise.

4. FALSE The same reason of the previous point.

5. TRUE If there is an elbow in the cumulated variance, the inclusion of the following principal components would not improve the representation of the data too much.