IE30301-Datamining Assignment 5 (70 Points)

Eldor Fozilov

29 May, 2022

Exercise 1

Determine the True/False (T/F) of the following statements and describe the reason(s). No point will be given without appropriate reason. You must also give your reasoning if you think the statement is true. [total 20 pts, 2 pts per each.]

1. For Multiple Linear Regression, We can construct T-test for testing overall model. False

   T-test is used for testing the statistical significance of a single variable in a model, not the overall model. For testing the overall, F-test can be used.

2. The parameters of logistic regression obtained by gradient descent algorithm and Newton-Raphson algorithm are same. True

   Since the cost function of logistic regression is convex, a local optimum is also a global optimum. Thus, it is very likely that gradient descent and Newton-Raphson algorithm converge at the sample place, thus giving identical parameters.

3. Among Linear Regression, PCA, Logistic Regression, LDA, Decision Tree, and KNN, only decision tree is non-parametric model. False

   KNN (the vanilla type) is also non-parametric.

4. The objective of Principal Component Analysis (PCA) is to derive uncorrelated features having the variance of the original data as maximum. True

   During the lecture about PCA, we learned that the goal of PCA analysis is to reduce the dimension of a data set and derive features that have zero correlation with each other and retain the variance of initial features as much as possible.

5. Linear Discriminant Analysis (LDA) will fail when the discriminatory information is not in the variance but rather in the mean of data. False

   Linear Discriminant Analysis should give good class assignments for sample records when the discriminatory information is not in the variance but rather in the mean of data as the objective function of LDA is designed to optimally address such a case. Linear Discriminant Analysis (LDA) will fail when the discriminatory information is <u>not in the mean</u> but <u>rather in the variance</u> of data.

6. In classification task, if there are imbalance of classes, we use accuracy rather than AUC of ROC Curve for evaluation method. False

   In classification task, if there are imbalance of classes (one class dominates over the other), we use true positive rate and false positive rate to access the performance of a model, rather than accuracy rate because accuracy rate does not account for the imbalance between classes and might give overly optimistic performance measure. Those two metrics (TPR and FPR) will be used as x-axis and y-axis variables (respectively) to plot the ROC curve for a range of probability thresholds and the performance of a model will be decided by looking at the AUC (area under the (ROC) curve).

7. To construct the contingency table for the result (output) of the soft classifier, a certain threshold needs to be predefined. True

   In soft classifiers, the output is in the form of a probability (between 0 and 1) and thus a threshold has to be predefined to assign classes to sample records.

8. In Bagging trees, individual weak learners are dependent of each other. False

   It is in boosting, not in bagging, that individual weak learners are dependent of each other.

9. K-Nearest Neighborhoods algorithm does more computation on classify time rather than train time. True

   Time complexity for training in KNN is O(1), which means that training time is constant and does not depend on sample size. However, time complexity for classifying new samples is O(n), which means the number of computations grows with number of training samples. Thus, KNN algorithm does more computation on classify time rather than train time.

10. For Support Vector Machine (SVM), deleting the support vectors will change the position of the hyperplane. True

    As hyperplane is determined by the support vectors, if those support vectors are deleted, then parameters w (vector) and b are recalculated, and those new parameters are going to be different. As a result, the position of the hyperplane will be different.

Exercise 2

Describe the "**main difference**" between the following two concepts of each question "**in one sentence.**" (note : you should answer only with the content discussed in class. Otherwise no points even though the answer is correct. You will also get no points when you answer differences that are not main.) [total 21 pts, 3 pts per each.]

1. **Predictive Data-Mining Task vs. Descriptive Data-Mining Task**

   Predictive data-mining task is associated with supervised machine learning methods and is generally referred to regression and classification tasks, where we try to predict a target variable using some feature variables, while **descriptive data-mining task** is

associated with unsupervised machine learning methods and is referred to tasks aimed at finding understandable patterns and underlying structure of the data.

2. **Principal Component Analysis vs. Linear Discriminant Analysis**

   LDA is a supervised machine learning method, which has the goal of finding new axis that will maximize separation between classes, while **PCA** is an unsupervised machine learning method, which has the goal of finding new axis that retain the most amount of initial data set's variance.

3. **Over-Fitting vs. Under-Fitting**

   Over-fitting occurs when a model performs extremely well on the training data, but performs poorly on the unseen test data, while **under-fitting** occurs when a model performs poorly on the training data and on the unseen test data.

4. **Support Vector(s) vs. Non-Support Vector(s)**

   Support vectors are those data points that are closest to the optimal hyperplane (they are located on the plus-plane or the minus-plane), while **non-support vectors** are those data points that are located outside of the region defined by plus-plane and the minus-plane.

5. **Hard Margin vs. Soft Margin**

   Hard margin is used when the classes are perfectly linearly separable, which means there will be no classification error of the training data, while **soft margin** is used when the classes are linearly non-separable, which means that there will be some classification errors of the training data (although the impact of noise will be reduced).

6. **Euclidean Distance vs. Pearson Linear Correlation**

   Euclidean distance is a similarity measure that focuses on how two expression values similar to each other in terms of their scales, while **Pearson linear correlation** is a similarity measure that decides how similar two expression profiles are based on the existence of linear co-movement between them.

7. **Single Linkage vs. Complete Linkage**

   Single linkage refers to the distance between two closest data points in separate clusters, while **complete linkage** refers to the distance between two farthest data points in separate clusters.

Exercise 3

Suppose we have a kernel function for arbitrary vectors $\mathbf{x}, \mathbf{z} \in \mathbb{R}^2$, which is defined as $K(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^T\mathbf{z})^2$. [14 pts]

3.1

We know that the given kernel function can be decomposed into the inner product of some transformation function $\phi(\cdot)$; $K(\mathbf{x}, \mathbf{z}) = \phi(\mathbf{x})^T\phi(\mathbf{z})$. For $\mathbf{x} = (x_1, x_2), \mathbf{z} = (z_1, z_2)$, find the transfor-

mation function $\phi(\cdot)$ [4 pts]

Answer:

$$K(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^T\mathbf{z})^2 = (x_1z_1 + x_2z_2)(x_1z_1 + x_2z_2) = x_1^2z_1^2 + 2x_1z_1x_2z_2 + x_2^2z_2^2 =$$
$$= \left[x_1^2,\ \sqrt{2}x_1x_2,\ x_2^2\right]^T \cdot \left[z_1^2,\ \sqrt{2}z_1z_2,\ z_2^2\right] = \phi(\mathbf{x})^T \cdot \phi(\mathbf{z})$$

3.2

For given vectors $\mathbf{x}_1 = (1, 2), \mathbf{x}_2 = (4, 3), \mathbf{x}_3 = (2, 0)$, Find the transformation of each vector using the transformation function you derived in **3.1** [3 pts]

Answer:

$$\phi(\mathbf{x_1}) = \begin{pmatrix} 1 \\ 2\sqrt{2} \\ 4 \end{pmatrix}, \quad \phi(\mathbf{x_2}) = \begin{pmatrix} 16 \\ 12\sqrt{2} \\ 9 \end{pmatrix}, \quad \phi(\mathbf{x_3}) = \begin{pmatrix} 4 \\ 0 \\ 0 \end{pmatrix}$$

3.3

Construct a kernel matrix K using $\mathbf{x}_1 = (1, 2), \mathbf{x}_2 = (4, 3), \mathbf{x}_3 = (2, 0)$. [4 pts]

Answer:

$$K = \begin{bmatrix} K(x_1, x_1) & K(x_1, x_2) & K(x_1, x_3) \\ K(x_2, x_1) & K(x_2, x_2) & K(x_2, x_3) \\ K(x_3, x_1) & K(x_3, x_2) & K(x_3, x_3) \end{bmatrix} = \begin{bmatrix} 25 & 100 & 4 \\ 100 & 625 & 64 \\ 4 & 64 & 16 \end{bmatrix}$$

3.4

Check if the kernel matrix you get in **3.3** is positive semi-definite or not. [3 pts] (To check positive semi-definiteness : you should verify it based on the fact that all eigenvalues of the given matrix are non-negative. https://www.symbolab.com/solver/matrix-eigenvalues-calculator/eigenvectors)

Answer:

After using the matrix-eigenvalues-calculator, we found the following:
$\lambda_1 \approx 2.89, \quad \lambda_2 \approx 15.45, \quad \lambda_3 \approx 647.68$
Since, all eigenvalues are non-negative, the kernel matrix is positive semi-definite.

Exercise 4

Suppose you are the teaching assistant(TA) of the Data Mining Course. The final exam is coming soon, and the professor asked you to make one question for the Final exam. As the TA, you wanted to check if the students have a solid understanding of the topic "SVM." Make a question that checks the concepts and theory you have learned on "SVM" with the below considerations. Provide the solution to your question as well. [15 pts]

1. The total score of the final exam is 100 points. The portion of this question you are making is 10 points.
2. The question should not be so easy nor so hard. You want the answer rate to be around 4 ~ 60%
3. Students who have studied the lecture note should be able to answer this question without much difficulty

To evaluate this question, all 5 TAs will grade this question together, and the median high score will be given as your score to this question. The grading will be based on

1. Completeness of the question. Is the question understandable and doable?
2. Does the question well satisfy the above considerations?
3. Does it well tests the understanding of the topic?

Question:

Let's say you are dealing with a data set where there are only two (balanced) classes, and there exist complex (nonlinear) relationships between them. On top of that, you are told that some "bad" people were able to add some noise (e.g., outliers) to the data set thinking that the performance of a machine learning model, which you are planning to implement for data classification, will show a big drop and they can exploit the failure of your model in some way. You know that SVM is a powerful classification model and you want to use it. However, what choice should you make in terms of kernel functions and hard-margin versus soft-margin implementations, so that the model shows a reasonable performance and does not let those "bad" people achieve what they want?

Answer:

Since we are dealing with a data set that has a complex structure, we can use non-linear SVM and to do that we transform the data using common kernels such as polynomial (of power 2) or Gaussian kernels. And since the data contain some noisy data points, we can use the soft-margin version of SVM to mitigate the effect of noise:

$$\Phi(\mathbf{w}) = \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + C\sum_{k=1}^{R}\xi_k$$

Note: R is the number of samples in the training set

Since the noise was not inherent in the data set, and was added by others, we want its effect to be as small as possible. Thus, we will choose very small (close to zero) value for the parameter C.