CS4220
Machine Learning

Written examination
02-02-2023, 13:30-16:30

- **Answer the questions on the answer sheets**
- **Don't forget to put your NAME and STUDENT NUMBER on the answer slides**
- **As much as possible, include the CALCULATIONS you made to get to an answer**
- **EXPLAIN how you come to your answer; the road to the answer is very important!**

1 Statements *(10 points)*

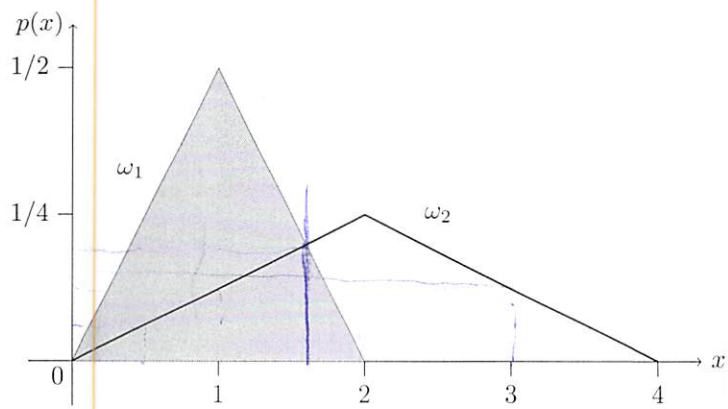
If a statement does not hold in general, but only under certain conditions that are not mentioned, then the statement should be marked as FALSE. 10 correct answers will give you 0 points; each additional correct answer gives you 1 point.

1. When the apparent error is smaller than the true error, we have overfitting.
2. The maximum rule for combining classifiers selects the classifier that overall performs best.
3. In expectation, the average 10-fold crossvalidation error is larger than the true error.
4. The Quadratic Discriminant classifier models each class as a Gaussian distribution.
5. The Bayes classifier is always a quadratic function.
6. The Fisher classifier models each class as a Gaussian distribution.
7. Single linkage clustering typically results in more compact clusters than average linkage clustering.
8. k-Means clustering results always reproduce exactly.
9. If no features are available, the optimal classifier is realized by assigning all objects to the class with the largest prior probability.
10. The k nearest neighbor classifier is sensitive to the rescaling of one of the features (if you have more than 1 feature).
11. The complexity of a classifier is directly related to the number of parameters that need to be estimated in the training phase.
12. In three dimensions, the nearest mean classifier can perfectly separate any configuration of four points from two classes

13. In hierarchical clustering, the largest jump in the fusion graph indicates where the dendogram should best be cut.
14. The Naive Bayes classifier models all the features as independent, given a class.
15. The classification error found for an infinite training set is for the Bayes classifier a non-increasing function of the number of features.
16. The logistic classifier models the log of the ratio of the class posterior probabilities as a linear function.
17. For a data set with 1,000-dimensional feature vectors, there are more than 1,000,000,000 different feature selection solutions.
18. The product rule for combining classifiers is based on the assumption that, for all individual classes, the base classifiers are estimated in independent feature spaces.
19. A reject option is helpful for constructing classifiers in the case when the cost of a reject is higher than the cost of an erroneous classification.
20. The neural network classifier minimizes the number of erroneously classified objects of the training set.

2 Classification and ROC curves *(12 points)*

Consider the following triangular pdf's of two classes in a 1D feature space:



a. Assume that the class prior probabilities are equal. Derive the posterior probabilities of the following objects: $x = 3$, $x = 0.5$, $x = 1$. To which classes are these objects therefore assigned? *(2 points)*

b. Where is the decision boundary of the Bayes classifier (assuming equal class priors)? *(2 points)*

c. How large is the Bayes error (assuming equal class priors)? *(2 points)*

Now assume that my classifier is very stupid: it only sets a threshold θ on feature x. If the feature value is larger than the threshold, the classifier predicts class ω_2, and otherwise it predicts class ω_1.

d. Assume that the threshold is set to $\theta = 1$. How large is the error on class ω_1 and on ω_2? Is the total error larger, equal, or smaller than the Bayes error? *(2 points)*

e. If the costs of misclassifying objects from class ω_1 to class ω_2 is 10 times the cost of misclassifying objects from class ω_2 to class ω_1, where should you put the threshold θ? *(2 points)*

f. Draw the ROC curve by varying the threshold θ of the classifier. *(2 points)*

3 Classifiers *(10 points)*

Let us consider a two-class classification problem, with labels $y_i = +1$ and $y_i = -1$. Assume that on a training set the following classifiers have been trained: (1) a logistic classifier (`loglc`),(2) a quadratic discriminant classifier (`qdc`) and (3) a support vector classifier (`svc`) with a polynomial kernel of degree 2.

a. Give for all the classifiers the mathematical formula that defines how a new point $\mathbf{z}$ is being classified (so the function should use a feature vector as input, and give a label as output). *(2 points)*

b. What is the criterion that is optimized for each of the classifiers during training/learning? What parameters are being optimized during that training? *(2 points)*

c. What are the free parameters in each of the classifiers that have to be set by the user? How can you set them in general? *(2 points)*

d. Explain how expensive it is to train the given classifiers; or in other words, given a very large dataset, what classifier is easiest to train, and what is the most computationally expensive one to train? *(2 points)*

e. Explain how expensive it is to *evaluate* a new test object by the given classifiers: what is the easiest classifier to apply, and what is the most expensive one? *(2 points)*

4 Clustering *(6 points)*

Given five points on the vertices of a 5D hypercube: $(0,0,0,0,0)$, $(0,0,0,1,0)$, $(0,1,0,0,0)$, $(0,0,1,0,0)$, and $(0,1,1,1,1)$.

a. Calculate the full distance matrix based on the standard Euclidean distance. *(1 points)*

b. Use complete linkage to cluster these five points and draw the associated dendrogram. Make sure that the drawing is clear and unambiguous! *(3 points)*

c. Say, instead of the Euclidean distance, we use the Hamming distance between the vectors. (The Hamming distance simply counts the number of positions in two vectors at which their entries differ.) How does the dendrogram change? *(2 points)*

5 Weighted Least Squares *(8 points)*

Let us consider a weighted version of standard least squares regression in a 1-dimensional feature space. In standard least squares, we minimize the total squared loss

$$L(a) = \sum_{i=1}^{N} (ax_i - y_i)^2 \tag{1}$$

on the N training data points (x_i, y_i) with x_i the 1-dimensional input and y_i the 1-dimensional output. In the setting we look at here, we now also have a weight $w_i > 0$ associated with every (x_i, y_i) and consider the weighted squared loss:

$$L(a) = \sum_{i=1}^{N} w_i (a x_i - y_i)^2. \tag{2}$$

Note that we do not consider an intercept.

a. Show, for the case of 1D inputs that we consider, that the minimizer equals

$$\hat{a} = \frac{\sum_{i=1}^{N} w_i x_i y_i}{\sum_{i=1}^{N} w_i x_i^2}.$$

(3 points)

b. Give the choices of weights for which the weighted least squares solution becomes equal to the minimizer of the standard least squares problem from Equation (1). *(2 points)*

c. Assume that the true input-output relation is given by $y = x + \varepsilon$, where ε is additive Gaussian noise with mean zero and standard deviation σ.

Furthermore, consider two datapoints that are far apart, namely $(x_1, y_1) = (1, 0.9)$ and $(x_2, y_2) = (10^{42}, 10^{42} + 10^{41})$. In other words, y_1 is 10% smaller and y_2 is 10% larger than x_1 and x_2, respectively. Will the standard least squares be biased to fit towards datapoint 1, datapoint 2 or is there no such tendency? Motivate your answer. *(3 points)*

6 Bayesian Networks *(8 points)*

Consider a joint probability distribution over 5 random variables A, B, C, D, E.

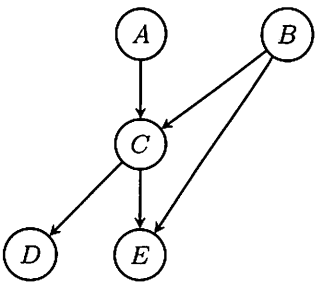
a. If each variable is a categorical variable with 3 possible values. In general, how many parameters do we need to describe this distribution? *(2 points)*

b. Suppose we can factorise the the distribution as follows:

$$P(A)P(B|A)P(C|A,B)P(D|A,B,C)P(E|A,B,C,D).$$

How many parameters do we need to describe the distribution now? *(2 points)*

c. Now, suppose a different factorization of the joint distribution, that is given by the following Bayesian network:



Which conditional independences hold? Mark all independences on the answer slide. *(4 points)*