Fondements de l'Apprentissage Machine (IFT 3395/6390)

Mid-term exam

Professor: Pascal Vincent

Thursday october 13^{th} 2016
Duration: 2h00

Allowed documentation: 2 two-sided sheet of paper (letter format 8" 1/2 x 11") with your own course summary.

First name:
Last name:
Code permanent:
IFT3395 or IFT6390:
Study program (and lab if nay):

Exam is 100pts. Write directly in the spaces left blank. Answers should be concise yet precise. **Good luck!**

Notation

For all questions, we suppose we are working with a starting dataset containing n examples denoted $D_n = \{z^{(1)}, \ldots, z^{(n)}\}$ with, in the supervised case, $z^{(k)} = (x^{(k)}, y^{(k)})$ where $x^{(k)} \in \mathbb{R}^d$ is the input and $y^{(k)}$ is the corresponding target. You must respect the notations specified within this exam (adapt known equations as needed).

1 Work situation (15 pts)

You are hired by a company that makes identity verification systems, and is developing a new face recognition system for a big client. The system must be capable of recognizing the faces of about twenty authorized people and distinguish them from any other non- authorized person. The company has a data base containing 200 000 labeled image faces (identified as authorized or non-authorized). A colleague of yours tells you that he applied 3 variants of classification algorithms, that he trained on the 200 000 images. The first yielded 4% classification error on the 200 000 images, the second 2%, and the third 0.3%. Since his experiment clearly shows that the third had a better performance, he wants to use it in the new system.

1. Do you agree with him? Explain/justify your answer.

2. If you disagree, how would you propose to decide which one of the variants should be used? In addition, the client requires a reliable estimate of the performance he can expect from the running system. How would you proceed? Please explain in details in your own words.

2 Over-fitting, under-fitting, capacity and model selection (15 pts)

Reminder: the "capacity" of a machine learning algorithm corresponds, informally, to the "size" or "richness" or "complexity" of the considered set of functions among which it searches for the best prediction function.

Answer T for True or F for False (or leave a blank) to the left of each of the following statements: +1 for a correct answer, -1 for a wrong answer, 0 for an abstention (It's thus better not to answer a question you are unsure about. Minimum for the exercise is 0/15, maximum is 15/15).

1. The more examples we have for training, the higher the risk of over-fitting.
2. A 1-nearest-neighbor classifier (1-NN) has a larger capacity than a $n-$nearest-neighbor classifier (k-NN with $k = n$).
3. The capacity of a learning algorithm can generally be controlled through the values of its *hyper-parameters*.
4. The larger the capacity of a machine learning algorithm, the better will its prediction be on new test examples.
5. The effective capacity of Parzen windows algorithms with a Gaussian kernel decreases as we increase the width of the kernel (the standard deviation of the Gaussians).
6. Over-fitting yields a low error rate on the validation set.
7. Under-fitting yields too large an error rate, both on the training set, and on the validation set.
8. When given the choice between several learning algorithms, we should choose the one that manages to best learn the examples on which it is trained.
9. The *Perceptron* algorithm has a higher capacity than the *1-Nearest-Neighbor* classifier.
10. For a fixed training set, generally the more parameters (scalars) there are to learn, the higher the risk of over-fitting.
11. The risk of under-fitting increases as we increase the capacity of the algorithm.
12. A Parzen windows classifier using a Gaussian kernel with too large a width σ will lead to over-fitting.
13. The larger the capacity of a machine learning algorithm, the fewer mistakes it will make on a complicated training set.
14. A machine learning algorithm with a larger capacity (than another) will tend to have a larger bias and a smaller variance.
15. If we were to choose the *hyper-parameter* values that yield the smallest error rate on the training set (on which an algorithm learns its *parameters*), it would always lead to choosing *hyper-parameter* values that yield the largest possible capacity.

3 Bayes Classifier and decision boundary (30 pts)

3.1 Naive Bayes Classifier

1. Given an input example $x \in \mathbb{R}^d$ that we wish to classify in one amongst m classes (variable Y). Provide a **detailed expression** of the class probability predicted by a **naive** Bayes classifier**.** Class-conditional probability densities are not supposed to be modeled by Gaussians in this question, Use $p()$ to denote these densities. In any case **explain/define your notations**. Note: we want a *detailed expression* (in which the naive hypothesis will be apparent).

   $P(Y = y|x) =$

2. What in this expression is "naive"? Why?

3.2 Bayes Classifier with Gaussian densities

1. Consider a classification problem with $m = 2$ classes (numbered 1,2). Consider a Bayes classifier where we modeled class-conditional densities each using an isotropic Gaussian. We restate the formula of the p.d.f. of an isotropic Gaussienne in d dimensions with mean μ and standard deviation $\sigma \in \mathbb{R}^+$:

$$\mathcal{N}_{\mu,\sigma^2}(x) \quad = \quad \frac{1}{(2\pi)^{\frac{d}{2}}\sigma^d} \exp\left(-\frac{1}{2}\frac{\|x-\mu\|^2}{\sigma^2}\right) \quad = \quad \frac{1}{(2\pi)^{\frac{d}{2}}\sigma^d} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{d}(x_i-\mu_i)^2\right)$$

   Given such a trained Bayes classifier, whose learned parameters are $\theta = \{\mu^{(1)}, \sigma^{(1)}, \pi^{(1)}, \mu^{(2)}, \sigma^{(2)}, \pi^{(2)}\}$ where $\pi^{(1)}$ and $\pi^{(2)}$ are the prior-probabilities of the two classes. Give a precise and detailed expression of the class probabilities predicted by this classifier:

   $P(Y = 1|x) \ =$

   $P(Y = 2|x) \ =$

2. Using your answer to the previous question, give a detailed expression that will define, for this specific classifier:

   (a) its decision boundary

   (b) the decision region of class 2

3. Now express the decision boundary using the *log* of the class probabilities rather than the class probabilities. From there, for the case where the 2 Gaussians are constrained to share the same standard deviation $\sigma^{(1)} = \sigma^{(2)} = \sigma$, show mathematically that the decision boundary corresponds to a hyper-plane.

4. In which specific case, does this classifier yield a decision function that is identical to the one obtained with a nearest mean (centroid) classifier ?

4 Histogram for regression (40 pts)

1. In a **regression** task, what is the nature of what the learned prediction function must predict?

2. Draw a graph showing an example of a regression data set containing $n = 20$ examples with inputs of dimension dimension $d = 1$, (Choose a set where examples are not all aligned along a straight line, and where inputs are regularly spaced).

3. Write the loss function (cost) L generally used for regression problems (clearly define/state what are the parameters of this function).

4. Provide a detailed expression of the empirical risk associated to this loss function, incurred by a predictor f_θ on a data set D (this detailed expression must not call upon L : you must have replaced these calls by their expression in the previous question).

5. Briefly explain in English, in your own words, what a "histogram" approach for **regression** consists of

6. Specify, for such a regression histogram, what would typically be:

   (a) its hyper-parameters

   (b) its parameters

7. Explain which hyper-parameter will serve to control the "capacity" of a histogram. For what kind of values of this hyper-parameter do we risk over-fitting? Why, what happens then?

8. For a regression histogram built with hyper-parameter values fixed by the user, explain in your own words how we would proceed to *estimate* its generalization error?

9. For this question, we suppose that the data set and hyper-parameters are chosen such that no histogram bin is empty. On the graph of sub-question 2, draw (also adding a legend to your graphic):

   (a) as a dashed line, the curve of the predictions obtained with a regression histogram whose choice of hyper-parameters yields too high a capacity

   (b) as a thin solid-line, the curve of the predictions obtained with a regression histogram whose choice of hyper-parameters yields too low a capacity and under-fitting

   (c) as a bold (thick) solid-line, the curve of the predictions obtained with a regression histogram whose choice of hyper-parameters seem most appropriate for your data.

10. For this question, we suppose that the data set and hyper-parameters are chosen such that no histogram bin is empty. Draw another graph with the shape of the "learning curves" (using a solid line for the training-set error; AND dashed line for the error on a validation set from the same distribution) associated to the search of this hyper-parameter for your histogram. Clearly label your axes.

11. Write a pseudo-code for the training function of a histogram for regression in dimension **d=3** (clearly name and define precisely any variable or parameter you use)

12. Write a pseudo-code for the prediction/use function of your trained histogram. In case the prediction function cannot predict anything, make it return 0 (not generally a recommended strategy, but here to simplify).

13. Write a pseudo-code to determine an appropriate value of hyper-parameters. Write is as a function that will allow training a regression histogram without the user having to specify any hyper-parameter.

14. Provide a mathematical expression (rather than an algorithmic pseudo-code) for the prediction function you wrote in subquestion 12.

15. We suppose hyper-parameter values have been specified by the user. *Express* the specific empirical risk minimization problem that would allow learning (training) the parameters of your histogram regressor (use your answers to points 4 and 14).

16. *Analytically solve* this optimization problem. What have you thus shown?