IFT6390 Fundamentals of Machine Learning

Professor: Ioannis Mitliagkas

Homework 1

- This homework is to be done in teams of 2 people. Make sure to have noted the name of all team members at the top of the report as well as at the top of the comments of each file you are submitting.
- It is required to hand in the report in electronic format (.pdf) All source code files you have created or adapted have to be submitted as well. The coding part is to be done in python (using the libraries numpy and matplotlib), and you can of course get inspired by what has been done in the lab sessions.
- You can submit your python code as an ipython notebook, .ipynb. In order to produce a report including mathematical formulas, you can use a software of your choice; LaTeX, LyX, Word; you can even write the theoretical parts directly into your iPython notebook (by typing the equations in MathJaX format for display). In any case you are asked to export your report which you will submit in .pdf format.
- The submission is to be done via StudiUM. There should be a single submission per team (one team member does the submission). Make sure to have noted the name of all team members at the top of the report. In case you have a lot of files to submit, you can also (if it is more practical) make an archive (.zip or .tar.gz.) and upload the archive file.

1 Small exercise on probabilities [10 points]

A few years ago, a study was carried out with doctors in the United States, in order to measure their "probabilistic intuition". It included the following question:

A percentage of 1.5% of women in their 40s who take a routine test (mammogram) have breast cancer. Among women that have breast cancer, there is a 87% chance that the test is positive. In women that do not have breast cancer, there is a probability of 9.6% that the test is positive.

A woman in her forties who has passed this routine test receives a positive test result. What is the probability that it is actually breast cancer?

A) more than 90%

B) between 70% and 90%

C) between 50% and 70%

D) between 30% and 50%

E) between 10% and 30%

F) less than 10%

95% of doctors surveyed responded B). What do you think? Formalize the question and calculate the exact probability.
Hint: use Bayes rule ...

2 Curse of dimensionality and geometric intuition in higher dimensions [20 points]

1. We consider a hyper-cube in dimension d (this is a generalization of the $2D$ square and the $3D$ cube) with side length c (which can be expressed in cm for example). What is the volume V of this hypercube?

2. We define a random vector X of dimension d ($x \in \mathbb{R}^d$) distributed uniformly within the hypercube (the probability density $p(x) = 0$ for all x outside the cube). What is the probability density function $p(x)$ for x inside the cube? Indicate which property(ies) of probability densities functions allow you to calculate this result.

3. Consider the outter shell (border) of the hypercube of width 3% of c (covering the part of the hypercube extending from the faces of the cube and $0.03c$ inwards). For example, if $c = 100$cm, the border will be 3cm (left, right, top, etc ...) and will delimit this way a second (inner) hypercube of side $100 - 3 - 3 = 94$cm. If we generate a point x

according to the previously defined probability distribution (by sampling), what is the probability that it falls in the border area? What is the probability that it falls in the smaller hypercube?

4. Numerically calculate the probability that x will fall in the narrow border for the following values of d: $1, 2, 3, 5, 10, 100, 1000$.
5. What do you conclude about the distribution of points in higher dimensions, which is contrary to our intuition in smaller dimensions?

3 Parametric Gaussian density estimation, v.s. Parzen window density estimation [35 points]

In this question we consider a dataset $D = \{x^{(1)}, \ldots, x^{(n)}\}$ with $x \in \mathbb{R}^d$.

1. Suppose we have trained the parameters of an **isotropic** Gaussian density function on D (by maximizing the likelihood) in order to estimate the probability density function.
   (a) Name these parameters and indicate their dimension.
   (b) If we learn these parameters using the principle of maximum likelihood estimation, express the formula which will give us the value of the optimal parameters as a function of the data points in D — indicate only the formula that calculates the result, you are not asked to rederive it (the formulas for the maximum likelihood estimator can be found at the end of slide set number 5 on the Gaussian distribution).
   (c) What is the algorithmic complexity of this training method, i.e. of the method calculating these parameters?
   (d) For a test point x, write the function that will give the probability density predicted at point x: $\hat{p}_{gauss-isotrop}(x) = ?$
   (e) What is the algorithmic complexity for calculating this prediction at each new point x?
2. Now consider that one uses Parzen windows with an isotropic Gaussian kernel of width (standard deviation) σ instead, and that these Parzen windows were trained on D.

(a) Suppose that the user has fixed σ. What does the "training/learning" phase of these Parzen windows consist of?

(b) For a test point x, write in a single detailed formula (i.e. with exponentials), the function that will give the probability density predicted at point x: $\hat{p}_{Parzen}(x) =?$

(c) What is the algorithmic complexity for calculating this prediction at each new point x?

3. Capacity/Expressivity

   (a) Which one of these two approaches (parametric Gaussian v.s. Parzen Gaussian kernel) has the highest *capacity* (in other words, higher expressivity)? Explain.

   (b) With which one of these approaches, and in which scenario, are we likely to be over-fitting (i.e. memorizing the noise in our data)?

   (c) The value σ in Parzen windows is usually treated as a hyper-parameter, whereas for parametric Gaussian density estimation it is usually treated as a parameter. Why?

4. Now consider parametric density estimation with a diagonal Gaussian density function.

   (a) Express the equation of a diagonal Gaussian density in $\mathbb{R}^d$. Specify what are its parameters and their dimensions.

   (b) Show that the components of a random vector following a diagonal Gaussian distribution are independent random variables.

   (c) Using $-\log p(x)$ as the loss, write down the equation corresponding to the empirical risk minimization on the training set D (in order to learn the parameters).

   (d) Solve this equation analytically in order to obtain the optimal parameters.

4 Practical part: density estimation [35 points]

1. Implement a diagonal Gaussian parametric density estimator. It will have to work for data of arbitrary dimension d. As seen in the labs, it should have a **train()** method to learn the parameters and a method **predict()** which calculates the log density.

2. Implement a Parzen density estimator with an isotropic Gaussian kernel. It will have to work for data of arbitrary dimension d. Likewise it should have a **train()** method and a **predict()** method that computes the log density.

3. 1D densities: From the Iris dataset examples, choose a subset corresponding to one of the classes (of your choice), and one of the characteristic features, so that we will be in dimension $d = 1$ and produce a single graph (using the plot function) including:

   (a) the data points of the subset (displayed on the x axis).

   (b) a plot of the density estimated by your parametric Gaussian estimator.

   (c) a plot of the density estimated by the Parzen estimator with a hyper-parameter σ (standard deviation) too small.

   (d) a plot of the density estimated by the Parzen estimator with the hyper-parameter σ being a little too big.

   (e) a plot of the density estimated by the Parzen estimator with the hyper-parameter σ that you consider more appropriate. Use a different color for each plot, and provide your graph with a clear legend.

   (f) Explain how you chose your hyper-parameter σ.

4. 2D densities: Now add a second characteristic feature of Iris, in order to have entries in $d = 2$ and produce 4 plots, each displaying the points of the subset of the data (with the plot function), and the contour lines of the density estimated (using the contour function):

   (a) by the diagonal Gaussian parametric estimator.

   (b) by the Parzen estimator with the hyper-parameter σ (standard deviation) being too small.

   (c) by the Parzen estimator with the hyper-parameter σ being a little too big.

   (d) by the Parzen estimator with the hyper-parameter σ that you consider more appropriate.

   (e) Explain how you chose your hyper-parameter σ