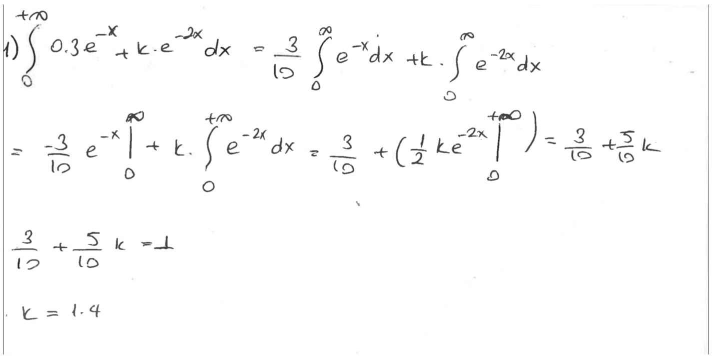
Machine Learning

Ufuk DOGAN

2018/11/02

1 Exercise 1: Theoretical Probability

1.1 First Part



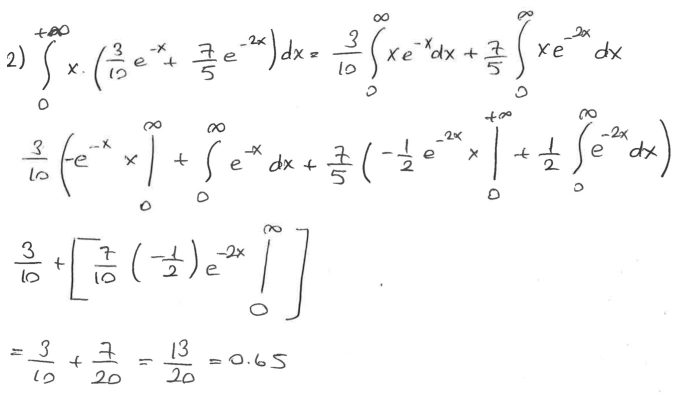
1) $\int_0^{+\infty} 0.3e^{-x} + k.e^{-2x} dx = \frac{3}{10}\int_0^{\infty} e^{-x}dx + k.\int_0^{\infty} e^{-2x}dx$

$= \frac{-3}{10} e^{-x}\Big|_0^{\infty} + k.\int_0^{+\infty} e^{-2x}dx = \frac{3}{10} + \left(\frac{1}{2} k e^{-2x}\Big|_0^{+\infty}\right) = \frac{3}{10}+\frac{5}{10}k$

$\frac{3}{10}+\frac{5}{10}k = 1$

$k = 1.4$

1.2 Second Part



2) $\int_0^{+\infty} x.\left(\frac{3}{10}e^{-x}+\frac{7}{5}e^{-2x}\right)dx = \frac{3}{10}\int_0^{\infty} xe^{-x}dx + \frac{7}{5}\int_0^{\infty} xe^{-2x}dx$

$\frac{3}{10}\left(-e^{-x}x\Big|_0^{\infty} + \int_0^{\infty} e^{-x}dx + \frac{7}{5}\left(-\frac{1}{2}e^{-2x}x\Big|_0^{+\infty} + \frac{1}{2}\int_0^{\infty} e^{-2x}dx\right)\right.$

$\frac{3}{10} + \left[\frac{7}{10}\left(\frac{-1}{2}\right)e^{-2x}\Big|_0^{\infty}\right]$

$= \frac{3}{10} + \frac{7}{20} = \frac{13}{20} = 0.65$

2 Exercise 2: Theory of SVMs

2.1 Explain in your own words how an SVM roughly works. Make sure to add key characteristics, advantages, and disadvantages.

The main idea of the support vector machines is, finding an optimal hyperplane which is the line that "best" seperates two classes of points, for linearly seperable patterns and also, extend the patterns that are not linearly seperable by transformations of original data to map into new space, with a function called Kernel.

The data set which consists of inputs and outputs will be the input of SVM functions. We will use the SVM function's output of data set specifying support vectors. Support vectors are the nearest data points to the hyperplane.

In hyperplane we have one minus plane, one plus plane and one classifier boundary.

Assume that we have 2 different outputs black and white. Some of the black results are so near to white ones and that's why linear classification is not working so efficient.

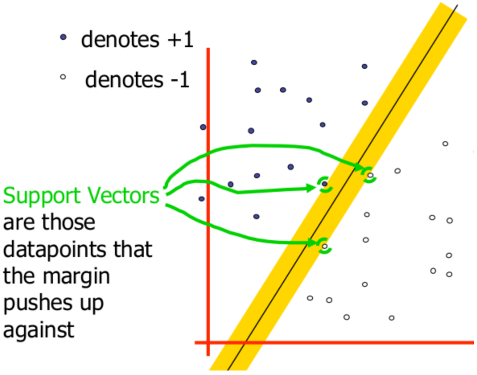
In our example, the closest black and white outputs are our support vectors and the space of between them is the our classifier boundary. These outputs are -1 and 1, let's assume that black ones are 1 and white ones are -1.

Our aim is, extending the area between minus plane and plus plane. By this way, we will have classify outputs easily.

Using SVM we try to find the optimal solution and it works very well, we can count this as an advantage.

The main disadvantage of using SVM is, it has too many important parameters which should be set so carefully.

Below, you can find the image too.


• denotes +1
◦ denotes -1
Support Vectors are those datapoints that the margin pushes up against

2.2 Explain the objective functions of hard-margin and soft-margin support vector machine training as well as the constraints of the corresponding optimization problems.

In soft and hard margin we try to extend the margin between support vectors.

In soft margin, we let our model to have some relaxation to few points. If we consider these our margin may reduce considerably and our decision boundary will be poorer, so instead of considering them as support vectors we consider them as error points and give certain penalty for them.

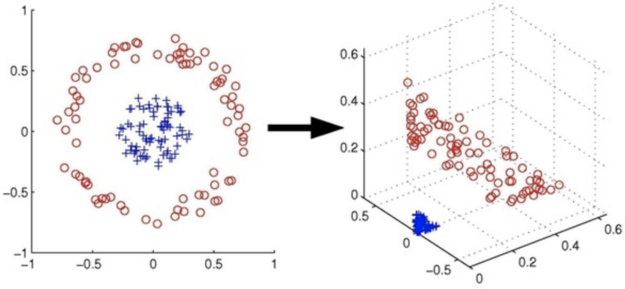
The allowance of softness in margins allows for errors to be made while fitting the model (support vectors) to the training/discovery data set.

Backwards, hard margins will result in fitting of a model that allows zero errors. Sometimes it can be helpful to allow errors in the training set, because it may produce a more generalizable model when applied to new datasets. Forcing rigid margins can result in a model that performs perfectly in the training set, but is possibly over-fit / less generalizable when applied to a new dataset.

2.3 What is the Kernel trick? When is a kernel valid? Provide necessary and/or sufficient conditions.

Kernel trick is using for classifying the data which is not linearly seperable.

The main idea of this trick is, we are taking the low dimensions data sets and converting them to the n dimensions using Kernel functions. Using the result we get in n dimensions, we can use the classify/solve the problem we have in lower dimensions.



As you can see, on the left side of the image we have a low dimension (2 dimension) data set and we can not seperate it linearly. Using Kernel functions, we can transform low dimension data set to the high dimension data set and at this point as you can see we can seperate it.

The kernel is valid if it's simmetric or a composition of valid kernels. Also, it should satisfy Mercer's Condition's too.

2.4 Consider the soft-margin optimization problem in point 2: write the equivalent dual problem in which the Kernel function appears.

$$max_\alpha \sum_{n=1}^{N} \alpha_n - 1/2 \sum_{n=1}^{N} \sum_{n=1}^{N} \alpha_1 \alpha_m y_1 y_m (\phi x_1 \cdot \phi x_m)$$

such that;

$$0 \leq \alpha_1 \leq C \quad \text{where} \quad \forall_n = 1, ..., N$$

and

$$\sum_{n=1}^{N} \alpha_n y_n = 0 \quad \text{where} \quad \forall_n = 1, ..., N$$

2.5 After solving the optimization problem, how would you classify a new point x? Express the formula in terms of the kernel function.

$$sign(w^T\phi(x) + b)$$

2.6 Explain which are the hyperparameters of an SVM and what are their effects

The hyperparameters of an SVM are; C (the soft margin constant), and any parameters the kernel function may depend on (width of a Gaussian kernel or degree of a polynomial kernel).

C (the soft margin constant) affects to the decision boundary. If we change our C with the lower variable, it allows to ignore points which are close the boundary and increase the the margin.

3 Exercise 3: Kernels I

- $K(x,y) = x^Ty + (x^Ty)^2$

  the result of this equation is a **valid kernel**, because we know that
  x^Ty

  is a linear vector and it is also a valid kernel. According to this information

  $(x^Ty)^2$

  equation is also valid kernel because it is a product of the x^Ty and if we do a sum operation between two valid kernel will have a valid kernel output.

- $K(x,y) = x^2e^{-y}, d = 1$

  the result of this equation is **not a valid kernel** because it is not symetric. We can understand if this equation's output is symetric or not, changing the places of the x and y.

  If we do it, we can see that the result of this two equations are different, so using this information we can say that this equation is not symetric, also is not a valid kernel.

- $K(x,y) = ck_1(x,y) + k_2(x,y)$, where $\mathrm{k}_1(x,y), k_2(x,y)$ are valid kernels in R^d

  The result of this equation is a **valid kernel if c is greater than zero** because the one condition of valid kernel is, it should be a gram matrix and this has to be positive.

4 Exercise 4: Kernels II

- **For a**

  We can see that we can not seperate the data set linearly so we can use **RBF or Polynomial Kernel Function**

- **For b**

  We can see that, this data set can seperable in multiple dimensions instead of 2 dimensions. So, we should use a kernel function which can mapping data from non linear separable space to high-dimensional separable space.

  As a reult of these, we can use **RBF or Polynomial Kernel Function**

- **For c**

  We can use the **Linear Kernel Function** because we can see that this data set can seperable linearly but in image we can see that some of the red and blue dots so near to each other. That's why we also use soft margin to use too.

- **For d**

  We can use the **RBF Kernel Function**. We can not use the Polynomial Kernel Function here because, it is too complex for Polynomial Kernel Function.

5 Exercise 5: SVMs

5.1 Suppose you have 2D examples: is the decision boundary of an SVM with linear kernel a straight line?

Yes, it will be a straight line.

5.2 Suppose that the input data are linearly separable. Will an SVM with linear kernel return the same parameters w regardless of the chosen regularization value C?

No. because if we increase C the result of the M is going to lower so, as a consequence of this maybe the support vectors will change and we will do missclassification.

5.3 Suppose you have 3D input examples (xi R3). What is the dimension of the decision boundary of the SVM with linear kernel?

It will be two dimension.

5.4 Is the computational effort for solving a kernel SVM increasing as the dimension of the basis functions increases? Why?

The computational effor will not change because we will use the Kernel Trick and with this way, we should not do any computation in high dimensions.

5.5 Suppose that after our computer works for an hour to fit an SVM on a large data set, we notice that x4, the feature vector for the fourth example, was recorded incorrectly, i.e., we use x 4 instead of x4 to train our model. However, your coworker notices that the pair (x 4, y4) did not turn out to be a support point in the original fit. He says there is no need to train again the SVM on the corrected data set, because changing the value of a non-support point can't possibly change the fit. Is this true or false?

No, it is false. Because maybe added new vector can be a new support vector