Midterm Revisions

CSCI-UA 9473, Introduction to ML

February 2014

1 Statistics and probability

Question 1.1. *In many frameworks, learning can be reduced to the estimation of the parameters defining a distribution from a few samples of that distribution. I.e given a distribution $p_{\boldsymbol{\theta}}(\boldsymbol{x}) = p(\boldsymbol{x}|\theta)$ with true parameters encoded by the vector $\boldsymbol{\theta} \in \mathbb{R}^d$, one wants to find an estimator $\hat{\boldsymbol{\theta}}$ of the parameters $\boldsymbol{\theta}$, relying on a small sample $\boldsymbol{x}_i, t_i$. Depending on whether we view the parameters $\boldsymbol{\theta}$ are fixed or as random, there are two main approaches that can be used to do parameter inference. List those approaches, explain how they differ and how they can be related through Bayes theorem.*

2 Linear Regression

Question 2.1. *Give the expression of the RSS model that is used in linear regression and explain each of the terms with a drawing.*

Question 2.2. *You are given a two classes dataset $\{\boldsymbol{x}_i, t_i\}$ for which you cannot get a proper geometric intuition. You decide to build feature vectors $\phi(\boldsymbol{x}_i)$ from your prototypes $\boldsymbol{x}_i$ that contain all the monomials up to some sufficiently large degree d.*

- *Assuming that you take a degree that is relatively large, how would you control the complexity of the model to avoid overfitting. Give 3 possible approches.*

- *The models from the first item, have all at least two parameters whose value now needs to be chosen carefully (the regularization parameter*

λ but also the maximum degree d). Explain how you can determine optimal values for those parameters. Give at least two approaches.

Question 2.3. *You are asked to find the parameters $\boldsymbol{\beta} \in \mathbb{R}^d$ and $\beta_0 \in \mathbb{R}$ that minimize the RSS criterion from Question 2.1 with ℓ_2 regularization. Write down the pseudo code of the gradient algorithm that you would use to find those parameters. You can assume a labeled dataset $\{\boldsymbol{x}_i, t_i\}$.*

Question 2.4. *Consider the following lines.*

```
from sklearn import linear_model
reg = linear_model.LinearRegression()
```

How would you learn the regression model for a dataset stored in variables `X,t`?

3 Linear Classification

Question 3.1. *Explain how you can use the RSS criterion to learn a classifier in the two classes framework.*

Question 3.2. *Give two straightforward extensions of the binary classifier to the multiclass problem. Explain how the prototypes are classified by each approach.*

Question 3.3. *When considering a classification problem with $K > 2$ classes, it is possible to extend the idea of the RSS criterion and to learn K separating planes simultaneously. This has the advantage of preventing class ambiguities such as those that are observed in simpler multiclass classification models. Explain how you can extend*

4 Kernels and SVMs

Question 4.1. *What is the main interest of Kernels? List a few kernels.*

Question 4.2. *Explain the Kernel trick. How can one turn the RSS model into a model relying only on the similarity between points? What is the final expression of the classifier and why is this expression interesting?*

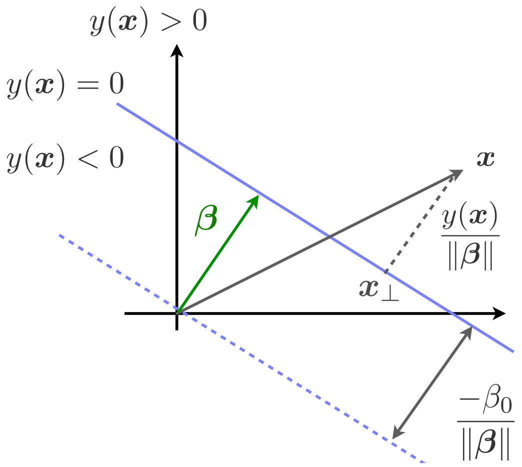

Figure 1: Exercise 4.3.

Question 4.3. *Consider Fig. 1. Explain this Fig and use it to give the expression of the distance of a point $\boldsymbol{x}$ to the plane $y(\boldsymbol{x}) = \boldsymbol{\beta}^T\boldsymbol{x} + \beta_0$.*

Question 4.4. *Consider Fig. 2. Explain this figure and give the mathematical expression of the corresponding classifier.*

5 Neural Networks

Question 5.1. *Give the mathematical expression of the perceptron and represent it with a diagram.*

Question 5.2. *Give the pseudo-code for the perceptron learning rule. This rule corresponds to minimizing a certain objective function. Give the expression of that objective. What can we say about the convergence of this algorithm?*

Question 5.3. *An extension of the simple perceptron, we consider the two hidden layers neural network shown in Fig. 3. The $\sigma(x)$ represent the activation functions and the $+1$ in magenta represent the bias terms which are multiplied by individual weights. Give the mathematical expression of this classifier. Given a dataset $\{\boldsymbol{x}_i, t_i\}_{i=1}^N$ with $t_i \in \{0, 1\}$ and $\boldsymbol{x}_i \in \mathbb{R}^2$, how would you learn the weights?*

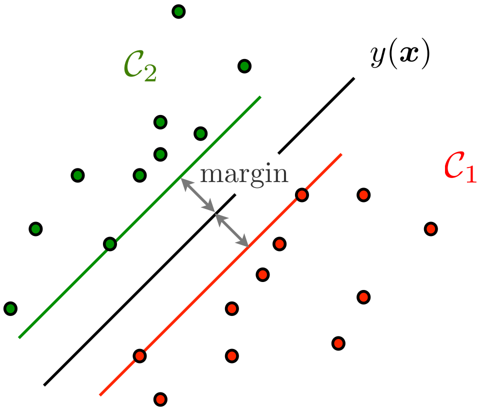

Figure 2: Exercise 4.4.

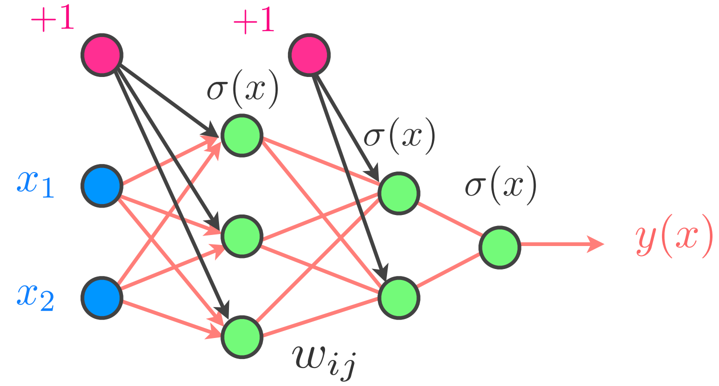

Figure 3: Exercise 5.3.

hidden_layer_sizes	tuple, length = n_layers - 2, default (100,)
activation	‘identity’,‘logistic’,‘tanh’
solver	‘lbfgs’, ‘sgd’, ‘adam’, default ‘adam’
alpha	float, optional, default 0.0001
learning_rate_init	double, optional, default 0.001
max_iter	int, optional, default 200

Table 1: Exercise 5.4 (Additional Specifications)

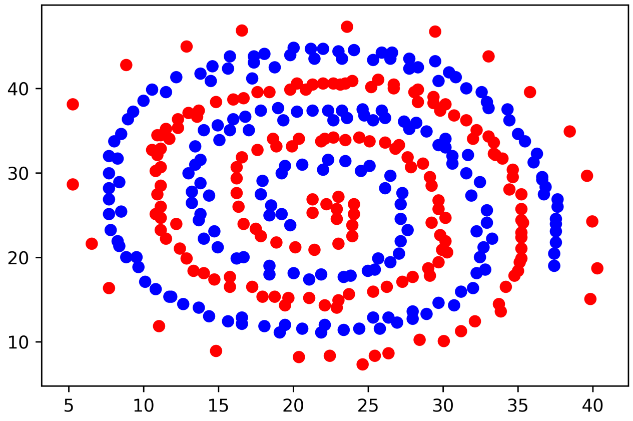

Figure 4: Exercise 5.4.

Question 5.4. *We want to design a multi-layer perceptron that can learn the dataset shown in Fig. 4. You are given the following lines in python.*

```
from sklearn.neural_network import MLPClassifier
```

Together with the documentation from Table 1:

Question 5.5. *Explain how Backpropagation can help derive the gradient efficiently in the training of neural networks.*