Name: Mouselinos Spyridon

Date: January 2020

Machine Learning and Computational Statistics

Homework 1

Exersize 1

- Name a classifier whose associate f(.) is of:
  - (A) parametric nature.
  - (B) non parametric nature.

Solution:

Let's think of the following scenario: We have 2 types of mushrooms that
look alike, however the first type are poisonous.
We need to classify them in two categories: 0 - poisonous, 1 - non-poisonous.
For this purpose we have decided to collect some data about specific features like their height and diameter.
By plotting a 2D plot of y-axis being the diameter and x-axis beign the height we
observe 2 clusters. The first may include mainly class 0 and the second mainly class 1.

A) A Parametric Classifier could be a linear separator with a fixed number of parameters - (n + 1) where n is the number of each entity's features - of those two classes, meaning the classifier would be a parametric function $f_\theta : R^2 \rightarrow R$ of a two dimentional input x = [height,diameter] and an one dimentional output y, that will "split" the classes appart.
If y >0 that could indicate class 1 and if y < 0 class 0.
The parameters $\theta = [\theta_0, \theta_1, \theta_2]$ of the linear function
$f(x) = \theta_0 + \theta_1 height + \theta_2 diameter$
would be calculated based on an optimization schema derived directly from the "fit" of our data or the "minimization" of a cost function. Such an example is the Logistic Regression Classifier.

B) A Non-Parametric Classifier of the above case would be a function f(.) that derives/adjusts/learns 0 parameters from the data, whereas uses the data themselves to derive to a result. Such a non-parametric appoach would be to classify a point as class 0 if the eucledian distance of it on the 2D plane is smaller towards its closest point in “class 0” rather than its closest point in “class 1”. Alternatively it would be classified as “class 1” if its eucledian distance between the closest “class 1” point is smaller than the distance between it and the closest “class 0” point. That is known as Nearest Neighbour Classifier and could be extended to take k “closest” neighbours into account. Thus only by using our data a kNearestNeighbour Classifier would consist a non-parametric approach to the same problem.

Exersize 2

- A) Define the parametric set of the quadratic functions $f_\theta : R \to R$ and give two instances of it. What is the dimensionality of $\boldsymbol{\theta}$?

Solution:

First of all we have that our input derives from R meaning that it is 1-D, $x = [x_1]$.
Furthermore our output derives from R meaning that it is also 1-D, $y = [y_1]$.
Quadratic Form in our f() means that we have a function of the form:
$y = \theta_0 + \theta_1 * x + \theta_2 * x^2$, thus our parameter vector would be: $\theta = [\theta_0, \theta_1, \theta_2] \in R^3$

Meaning that $F_{quadratic} = [f_\theta(.) : \theta \in R^3]$

Finally two instances are:
$f_\theta(x) = 1 + 2x + 3x^2, \theta = [1, 2, 3]$
$f_\theta(x) = 4x^2, \theta = [0, 0, 4]$

- B) Define the parametric set of the 3rd degree polynomials $f_\theta : R^2 \to R$ and give two instances of it. What is the dimensionality of $\boldsymbol{\theta}$?

Solution:

First of all we have that our input derives from R^2 meaning that it is 2-D, $x = [x_1, x_2]$.
Furthermore our output derives from R meaning that it is 1-D, $y = [y_1]$.

3rd Degree Polynomials in our f() means that we have a function of the form:
$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_2 x_1 + \theta_4 x_1^2 + \theta_5 x_2^2 + \theta_6 x_1^2 x_2 + \theta_7 x_2^2 x_1 + \theta_8 x_1^3 + \theta_9 x_2^3$,
thus our parameter vector would be: $\theta = [\theta_0, \theta_1, \theta_2, \theta_3, \theta_4, \theta_5, \theta_6, \theta_7, \theta_8, \theta_9] \in R^{10}$

Meaning that $F_{3deg} = [f_\theta(.) : \theta \in R^{10}]$

Finally two instances are:
$f_\theta(x) = x_1 + 2x_2 + 3x_2 x_1 + 4x_1^2 + 5x_2^2 + 6x_1^2 x_2 + 7x_2^2 x_1 + 8x_1^3 + 9x_2^3$,
$\theta = [0, 1, 2, 3, 4, 5, 6, 7, 8, 9]$

$f_\theta(x) = 0$,
$\theta = [0, 0, 0, 0, 0, 0, 0, 0, 0, 0]$

- C) Define the parametric set of the 3rd degree polynomials $f_\theta : R^3 \to R$ and give two instances of it. What is the dimensionality of $\boldsymbol{\theta}$?

Solution:

First of all we have that our input derives from R^3 meaning that it is 3-D, $x = [x_1, x_2, x_3]$.
Furthermore our output derives from R meaning that it is 1-D, $y = [y_1]$.
3rd Degree Polynomials in our f() means that we have a function of the form:

$f_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_2 x_1 + \theta_5 x_3 x_1 + \theta_6 x_3 x_2 + \theta_7 x_1^2 + \theta_8 x_2^2 + \theta_9 x_3^2 + \theta_{10} x_1^2 x_2 + \theta_{11} x_1^2 x_3 + \theta_{12} x_2^2 x_1 + \theta_{13} x_2^2 x_3 + \theta_{14} x_3^2 x_1 + \theta_{15} x_3^2 x_2 + \theta_{16} x_1 x_2 x_3 + \theta_{17} x_1^3 + \theta_{18} x_2^3 + \theta_{19} x_3^3$,

thus our parameter vector would be: $\theta \in R^{20}$

Meaning that $F_{3deg} = [f_\theta(.) : \theta \in R^{20}]$

Finally two instances are:
$f_\theta(x) = x_1 + x_2$
$\theta = [0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]$

$f_\theta(x) = 1$
$\theta = [1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]$

- D) Consider the function $f_\theta : R^5 \rightarrow R$,
  $f_\theta = \frac{1}{1+e^(-\theta^T x)}$. Define the associated parametric set and give two instances of it. What is the dimensionality of $\boldsymbol{\theta}$?

Solution:

First of all we have that our input derives from R^5 meaning that it is 5-D,
$x = [x_1, x_2, x_3, x_4, x_5]$.
Furthermore our output derives from R meaning that it is 1-D, $y = [y_1]$.
Here the function of y is not to be assumed, rather given in its form.
So Let :
$\theta^T = [\theta_0, \theta_1, \theta_2, \theta_3, \theta_4, \theta_5] \in R^6$
and $x = [1, x_1, x_2, x_3, x_4, x_5]$
then, $z = \theta^T x = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4 + \theta_5 x_5$ --> a scalar.
And we can write that: $f_\theta(x) = \frac{1}{1+e^(-z)} \in R$

Meaning that $F_{sigmoid} = [f_\theta(.) : \theta \in R^6]$

Finally two instances are:
$f_\theta(x) = 0.5$
$\theta = [0, 0, 0, 0, 0, 0]$

$f_\theta(x) = \frac{1}{1+e^{-2020}}$
$\theta = [2020, 0, 0, 0, 0, 0]$

- E) In which of the above cases f_θ is linear with respect to $\boldsymbol{\theta}$?

Solution:

```
In cases A,B,C
```

Exersize 3

- Verify that for two l-dimensional column vectors $\theta = [\theta_1, \theta_2, ..., \theta_l]^T$
  and $x = [x_1, x_2, \ldots, x_l]^T$ it holds: $(\theta^T x)x = (xx^T)\theta$.

Solution:

Let's see the first part of our equation: $(\theta^T x)x$ it is a multiplication of the form:
$[1 \times l] * [l \times 1] * [l \times 1] = [1 \times 1] * [l \times 1] = scalar * [l \times 1] = [l \times 1]\ vector$. From this we have that the first product is a scalar.
For a scalar z it holds that $z^T = z$ in our case: $(\theta^T x) = (\theta^T x)^T = (x^T \theta)$

We can thus transform the first part of our equation:
$(\theta^T x)x = (x^T \theta)x = x(x^T \theta)$we can alter the ordering of the multiplication because the $(x^T \theta)$ is a scalar.

However, due to the associative property of matrix multiplication we can re-write the last part as follows:

$x(x^T \theta) = (xx^T)\theta$ proving that finally:
$(\theta^T x)x = (xx^T)\theta$

Numeric Verification:
$\theta = [1, 2]^T$
$x = [0, 1]^T$

$(\theta^T x)x = 2 * [0, 1]^T = [0, 2]^T$
$(xx^T)\theta = [0, 0; 0, 1] * [1, 2]^T = [0, 2]^T$

Exersize 4

- Consider the vectors $x_n = [x_{n1}, x_{n2}, ..., x_{nl}]^T$ n = 1,...,N. Define the Nxl matrix X and N-dimensional column vector Y.
  Verify the following identities: $X^T X = \sum_{n=1}^{N} x_n x_n^T$ and $X^T Y = \sum_{n=1}^{N} y_n x_n$

Solution:

Let us define $\mathbf{x_n} = [x_{n1}, x_{n2}, ..., x_{nl}]^T = \begin{bmatrix} x_{n1} \\ x_{n2} \\ ... \\ x_{nl} \end{bmatrix}$ as an $l \times 1$ matrix.

Then $\mathbf{x_n^T} = [x_{n1}, x_{n2}, ..., x_{nl}]$ as an $1 \times l$ matrix.

Now the X matrix would be of the following form:

$$X = \begin{bmatrix} \mathbf{x_1}^\mathbf{T} \\ \mathbf{x_2}^\mathbf{T} \\ ... \\ \mathbf{x_n}^\mathbf{T} \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & ... & x_{1l} \\ x_{21} & x_{22} & ... & x_{2l} \\ ... & & & \\ x_{n1} & x_{n2} & ... & x_{nl} \end{bmatrix}$$ an $n \times l$ matrix.

And the matrix X^T would be of the following form:

$$X^T = \begin{bmatrix} \mathbf{x_1} & \mathbf{x_2} & ... & \mathbf{x_n} \end{bmatrix} = \begin{bmatrix} x_{11} & x_{21} & ... & x_{n1} \\ x_{12} & x_{22} & ... & x_{n2} \\ ... & & & \\ x_{1l} & x_{2l} & ... & x_{nl} \end{bmatrix}$$ an $l \times n$ matrix.

Now the product $X^T X = \begin{bmatrix} \mathbf{x_1} & \mathbf{x_2} & ... & \mathbf{x_n} \end{bmatrix} \begin{bmatrix} \mathbf{x_1}^\mathbf{T} \\ \mathbf{x_2}^\mathbf{T} \\ ... \\ \mathbf{x_n}^\mathbf{T} \end{bmatrix} =$

$= \mathbf{x_1 x_1}^\mathbf{T} + \mathbf{x_2 x_2}^\mathbf{T} + ... + \mathbf{x_n x_n}^\mathbf{T} =$

$= \sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}^\mathbf{T}{}_n$

Now regarding $X^T Y$:

X^T is a $l \times n$ matrix where Y is a $n \times 1$ matrix.

$$X^T Y = \begin{bmatrix} \mathbf{x_1} & \mathbf{x_2} & ... & \mathbf{x_n} \end{bmatrix} \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ ... \\ \mathbf{y}_N \end{bmatrix} =$$

$= \sum_{n=1}^{N} \mathbf{x}_n \mathbf{y}_n$

Exercise 5

- Write explicitly the derivation of the Least square estimator, following the line of proof given in the slides of the 1st lecture.

Solution:

Let $X = \{(x_n, y_n), x_n \in R^l, y_n \in R, n = 1, \ldots, N\}$ a known dataset of size N,
that consists of an array of $l - dimensional$ input vectors x, and a $single\ dimensional$ output vectors y.

Let also a parametric set of functions f(·) and a parameter vector θ
such as $F := f_\theta(\cdot) : \theta \in A \subseteq R^K$.

Assuming the mean square error loss function $L(y_n, f_\theta(x_n)) = (y_n - f_\theta(x_n))^2$ per point, we get the total error, aka cost function as follows:
$J(\theta) = \sum_{n=1}^{N} L(y_n, f_\theta(x_n))$

Now we aim to find a θ in order to minimize $J(\theta)$.

Step 1: *Let's define the per point relationship between y,x and θ.*
For each entry in our dataset we have that:
$y_i = \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + ...\theta_l x_{il} = \theta^{\mathrm{T}} x_i$

So let's define $\Theta = [\theta_0, \theta_1, ..., \theta_l]^T$ as the $(l+1) \times 1$ vector of parameters,

and $\mathbf{x_i} = [1, x_{i1}, x_{i2}, ..., x_{il}]^T$ as the $(l+1) \times 1$ vector of input i.

That transforms the upper equation to the following matrix form for all (x_i, y_i) pairs in the dataset:
$Y = \mathrm{X}\Theta$ where:

- Y is the $N \times 1$ matrix of outputs,
- X is the $N \times (l+1)$ matrix of inputs,
- Θ is the $(l+1) \times 1$ matrix of inputs,

Step 2: *Let's solve it under the MSE cost function.*

In matrix notation the Cost function $J(\theta) = \sum_{n=1}^{N} L(y_n, f_\theta(x_n))$ can be rewritten simply as:

$$J(\theta) = (X\Theta - Y)^2 = (X\Theta - Y)^T(X\Theta - Y) = ((X\Theta)^T - Y^T)(X\Theta - Y) =$$

$$(\Theta^T \mathrm{X}^\mathrm{T} - Y^\mathrm{T})(X\Theta - Y) = \Theta^T \mathrm{X}^\mathrm{T} X\Theta - \Theta^T \mathrm{X}^\mathrm{T} Y - Y^\mathrm{T} X\Theta + Y^\mathrm{T} Y$$

However in order to minimize we have to find the derivate w.r.t Θ and set it to 0.

Step 3: *Find minimum.*
Differentiating w.r.t Θ gives us:
$\nabla_\theta J(\theta) = \nabla_\theta \Theta^T \mathrm{X}^\mathrm{T} X\Theta - \nabla_\theta \Theta^T \mathrm{X}^\mathrm{T} Y - \nabla_\theta Y^\mathrm{T} X\Theta + 0 =$
$2X^T X\Theta - \mathrm{X}^\mathrm{T} Y - X^\mathrm{T} Y = 2X^T X\Theta - 2\mathrm{X}^\mathrm{T} Y$

Setting $\nabla_\theta J(\theta) = 0$ we get:

$2X^T X\Theta_0 - 2\mathrm{X}^\mathrm{T} Y = 0$
$X^T X\Theta_0 = \mathrm{X}^\mathrm{T} Y$
$\Theta_0 = (X^T X)^{-1} \mathrm{X}^\mathrm{T} Y$ provided the matrix $(X^T X)$ is invertible.

Exercise 6

- A) Estimate the initial velocity and the acceleration of the body, based on the above measurements, utilizing the least squares error criterion.

Solution:

We have a pair of 5 entities in our dataset, giving us $N = 5$.
Our parametric model used to approximate the unkown mechanism
uses 1 feature, meaning $l = 1$.
So far we can assume that we have $\theta = [\theta_0, \theta_1] \in R^2$
and $x_n = [1, x_1]$ for $n \in [1, N] = [1, 5]$.

We can formulate the following:

$$X = \begin{bmatrix} \mathbf{x_1}^\mathrm{T} \\ \mathbf{x_2}^\mathrm{T} \\ \mathbf{x_3}^\mathrm{T} \\ \mathbf{x_4}^\mathrm{T} \\ \mathbf{x_5}^\mathrm{T} \end{bmatrix} = \begin{bmatrix} 1 & \mathbf{1} \\ 1 & \mathbf{2} \\ 1 & \mathbf{3} \\ 1 & \mathbf{4} \\ 1 & \mathbf{5} \end{bmatrix}$$

$$Y = \begin{bmatrix} 5.1 \\ 6.8 \\ 9.2 \\ 10.9 \\ 13.1 \end{bmatrix}$$

and calculating the required matrices:

$$X^T X = \begin{bmatrix} 5 & 15 \\ 15 & 55 \end{bmatrix}$$

$$X^T Y = \begin{bmatrix} 45.1 \\ 155.4 \end{bmatrix}$$

So we have that: $(X^T X)\theta = X^T Y$ and because the matrix $(X^T X)$ has det = $275 - 225 = 50 \neq 0$ it is invertible. So we finally get :

$$(X^T X)^{-1} = \begin{bmatrix} 1.1 & -0.3 \\ -0.3 & 0.1 \end{bmatrix}$$

$$\theta = (X^T X)^{-1} X^T Y$$

Resulting in:
$$\begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix} = \begin{bmatrix} 2.99 \\ 2.01 \end{bmatrix}$$

So the Least Square Approximation of our function is v = 2.99 + 2.01*t

- B) Estimate the velocity of the body at t=2.3.

By setting t = 2.3 we get v = 7.613