Machine Learning - Problem set 1

BOISSIN Thibaut

February 2018

table of contents

I. Q1

1.

For this question we need to reconsider the model to be probabilistic : instead of comparing to a dataset (x_i, d_i) we need to compare to a function modelling the probability to have one or the other classes. This function will be noted f and as our algorithm gives deterministic results, we need to compare to a function that can be defined as :

$$f\prime(x_i) = \frac{\mid (x_i, 1) \mid}{\mid (x_i, 1) \mid + \mid (x_i, 0) \mid}$$

Which allow comparision.

We need then to determine $P(D \mid h)$ using it's definition first. Then we notice that we can use bayes rule as x is independent of h :

$$P(D \mid h) = \prod_i P(x_i, d_i \mid h) = \prod_i P(d_i \mid h, x_i) P(x_i)$$

In the previous expression, we can note $P(d_i \mid h, x_i)$ follows a Bernouilli distribution with $p = h(x_i)$. We get then $P(d_i \mid h, x_i) = h(x_i)^{d_i}(1 - h(x_i))^{1-d_i}$
Now we have an expression of $P(D \mid h)$ we can define h_{ML} with the argmax over h :

$$P(D \mid h) = argmax_{h \in H} \prod_{i=1}^{m} h(x_i)^{d_i}(1 - h(x_i))^{1-d_i} P(X_i)$$

As the last term of the product is independent of h we can drop it from the expression.

$$P(D \mid h) = argmax_{h \in H} \prod_{i=1}^{m} h(x_i)^{d_i}(1 - h(x_i))^{1-d_i}$$

Note : we can apply a log over his expression to simplify the search, as log is increasing and monotonic.

2.

By considering the case of a deterministic function with a zero-mean gaussian noise, we're actually approaching By looking at the way we constructed the error function we can notice that if the number of samples is great, the binomial law of the errors can be approached by a gaussian. This means that if m is very high, the results would be the same. But this needs to have more that 30 sample for each X_i. If we have exact probabilities, the results would also be the same.

II. Q2

The following figure (1) shows that the line that separate the points have the equation $X_b = X_a - 0.5$. By rearranging the equation we can choose $w_a = -1$, $w_b = 1$, and $\theta = -0.5$.

We can now infer build the table of Xor using A and not B :

For now we can apply the perceptron learning rule to get the weights, but we can also set these by hand. If you start with a perceptron doing the *or* function and adjust the

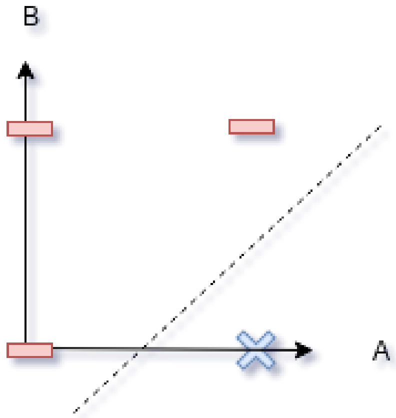

FIGURE 1 – linear separation of $A \wedge \neg B$

A	B	$A \wedge \neg B$	A xor B
F	F	F	F
F	T	F	T
T	F	T	T
T	T	F	F

TABLE 1 – Caption

weights to correct the last result (A = T and B = T). This can be done by decreasing the weight on A such as the $w_a A + w_b B < w_b B$ we then correct the result of the $3^r d$ row by putting a high weight on $A \wedge \neg B$.
The obtained set of weights is :

- $w_a = -0.5$
- $w_b = 0.5$
- $w_{a\&-b} = 1$
- $\theta = 0.2$ anything between 0 and 0.5 would work

III. Q4

The most common way to perform a regression task with a decision tree follow this process :

1. put the continuous feature into bins. Various strategies may be applied, for instance we can compute bins of similar cardinality over the training set (we hope this approaches real data distribution)
2. train the trees using bins as label
3. compute a function on each leaf that fits the training data for this leaf. For instance mean or meadian ca be used, but other types of function can be used (linear or quadratic regression for instance)

IV. Q6

If the data are linearly separable, the use of the KNN would be prefered. This is the case because the decision tree model have difficulties to mimic the behaviour of lines as conditions on the nodes are defined on one variable at a time. At the opposite, KNN, with a reasonably high value for K leads to a behaviour close to the SVM, which would fit nicely the data.

V. Q7

1.

First we may ask "how many parameters do we need to define this object ?". As we need only one parameter (the radius) We can then assume that it's VC dimension is 2 (number of paramter + 1). This is coherent with the defintion as it is possible to separate 2 points with a circle, but it's not possible to separate 3 point with a circle without moving the center.

2.

By applying the same strategy on the 3D sphere we note that we still can define it with a single param (the radius) as it is origin centered. So the VC dimension of the origin centered 3d sphere is 2.