MAC0460 – Introduction to Machine Learning

DCC / IME-USP — 2021

Questions regarding classes 01 to 05

This is a list of questions/topics to serve as a reference for the discussions in the next Monday's online class (May 3rd)

The perceptron algorithm

An interesting question that arose is the following: suppose $\mathbf{x}$ is such that $\text{sign}(\mathbf{w}^T\mathbf{x}) \neq \text{sign}(y)$. After updating the weight vector $\mathbf{w}$ as $\mathbf{w} \leftarrow \mathbf{w} + y\,\mathbf{x}$, can we say that we will have $\text{sign}(\mathbf{w}^T\mathbf{x}) = \text{sign}(y)$?

Sometimes we see variants of the perceptron algorithm. One of them performs weight updates considering a factor $\eta > 0$: $\mathbf{w} \leftarrow \mathbf{w} + \eta y\,\mathbf{x}$. What would be the utility of factor η ?

A perceptron just partitions the space in two halves. If you were allowed to use multiple perceptrons, would you be able to build arbitrary partitions ?

Linear regression

We used a cost function based on square root error ($(\mathbf{w}^T\mathbf{x} - y)^2$) to fit a hyperplane. Suppose now we would like to use the absolute error ($|\mathbf{w}^T\mathbf{x} - y|$). How do we optimize the cost function ?

Hyperplanes are simple and nice, but your data is often much more complex than something a hyperplane could explain. Could you fit polynomials instead lines? How do we do a polinomial regression ?

Logistic regression

Prove that $\dfrac{1}{1+e^{-z}} = \dfrac{e^z}{e^z+1}$

Prove that $1 - \theta(z) = \theta(-z)$

What is the shape of decision boundaries produced by the logistic regression algorithm ? Explain.

Logistic regression produces a "score" between 0 and 1 for each $\mathbf{x}$. What happens with this score as you get closer to the decision boundary ?

When we apply the gradient descent technique to the linear regression problem (with the squared difference) the amount the weight vector is updated is given by $\Delta w_j(r) = \sum_{n=1}^{N}(y^{(n)} - \hat{y}^{(n)})\mathbf{x}_j^{(n)}$. When we do the same with respect to the cross-entropy cost function (slide 32 of class 05), we also have $\Delta w_j(r) = \sum_{n=1}^{N}(y^{(n)} - \hat{y}^{(n)})\mathbf{x}_j^{(n)}$. Although visually they are the same, there is an important difference. What is the meaning of $\hat{y}^{(n)}$ in each case ?

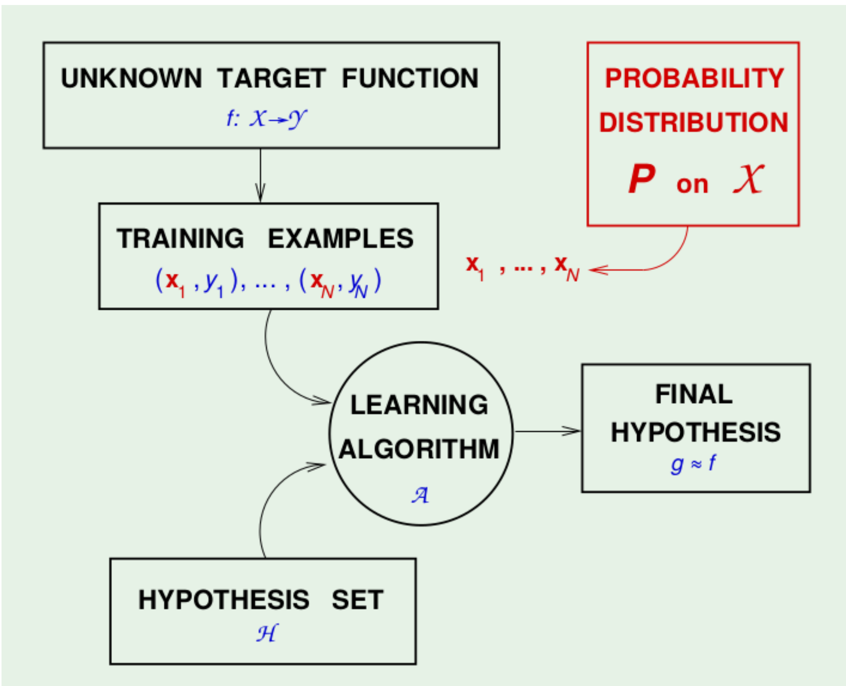
Suppose you have two binary classification problems. When you employ the logistic regression algorithm, for one of them, you get 98% of correct classification and for the other one you get 72% of correct classification. What can you say about your two datasets, based on these results ?

Other "random" topics

The three algorithms above are "linear models". What does that mean ? In classification problems, which type of decision boundaries are found ?

How can we perform multi-class classification (more than two classes) ?

Which are the components that you can "control" in the learning model diagram (see below) ?



Is there any computational problem that you have faced before and that could be solved using a machine learning based approach ? Describe the problem and outline the details of the approach.