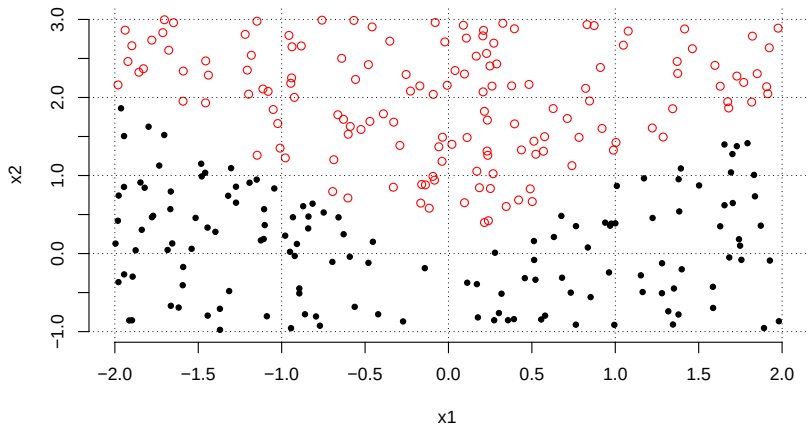
MSc HDA & ML: Machine Learning

Warmup for Lecture 6 - 17 Feb 2020

Spring 2020

Question 1

Here is some data for a two-class classification problem:



- What is the maximum margin classifier if you are restricted to linear decision boundaries, assuming a soft margin? What if you are no longer restricted to linear decision boundaries? (Draw your answers above.)
- Can you come up with a feature transformation ϕ to enable a linear decision boundary in the feature transformed space? ϕ should take a two-dimensional vector as input and it can return any dimensionality, e.g. here are two possibilities:

$$\phi_1\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right) = x_1 \cdot x_2$$

$$\phi_2\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right) = \begin{pmatrix} x_1 \\ |x_2 - x_1| \\ x_1 \cdot x_2 \end{pmatrix}$$

- Let us assume you choose a feature transformation $\phi : \mathcal{R}^2 \to \mathcal{R}$:

$$\phi\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right) = x_1 + x_2^2$$

In the original space $\mathcal{R}^2$, we have two vectors: $a = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$, $b = \begin{pmatrix} -4 \\ 2 \end{pmatrix}$.

Considering them as points (i.e. placing the tail of the vectors at the origin), what is the distance between them?

In the transformed space $\mathcal{R}$ what is the distance between them?

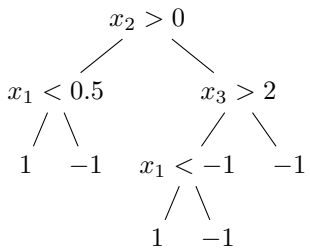
In the original space, what is the dot product $\langle a, b \rangle$? What is the dot product in the transformed space, $\langle \phi(a), \phi(b) \rangle$?

Write down the generic form of the kernel $k\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}\right)$ that calculates this dot product, i.e.:

$$k\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}\right) = \langle \phi\left(\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right), \phi\left(\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}\right) \rangle = ?$$

Question 2

We will cover decision trees today. Decision trees are a highly interpretable method, which means they are very popular in settings involving non-expert users (e.g. developing a decision support system for use by physicians and nurses in a hospital). Consider a dataset with three inputs x_1, x_2, and x_3 and a binary (positive/negative) output y. Here is a decision tree—you check a condition, e.g. $x_2 > 0$ and if it is satisfied, you follow the left branch. Otherwise, you follow the right branch. When you reach a so-called "leaf" node, you have your prediction, either 1 or -1.



The one bit of mathematics we will need to develop decision trees is a splitting rule. Given training data, how did we decide that the tree should split using variable x_2 at the root? The answer is that the splitting rule minimized a quantity called "impurity" or equivalently, it maximised the "information gain". There are different ways of defining impurity and information gain.

Let us develop the intuition for the **Gini index** (a measure of impurity) for a given dataset consisting of a positive cases and b negative cases as follows. If we used the base rate to make our classifications, that is, $\frac{a}{a+b}$ of the time we predict "positive" and $\frac{b}{a+b}$ of the time we predict "negative," what would we expect the misclassification rate to be? (Try to work this out yourself before reading on.)

We can calculate the misclassification rate as follows:

$$P(\text{positive case}) \times P(\text{we predict negative}) + P(\text{negative case}) \times P(\text{we predict positive})$$

Substitute the probabilities above into this expression to find:

$$\text{Gini index} = 2\frac{a}{a+b} \cdot \frac{b}{a+b} \tag{1}$$

Now define $p = \frac{a}{a+b}$ to be the probability of a positive case in the data. Calculate the Gini index in terms of p. Make a plot of the Gini index, varying p from 0 to 1.

What value of p gives the largest Gini index? Gini index is one way of measuring impurity; can you see why this is called "impurity"?

We can generalise the Gini index to multiple classes using the logic above. Given K classes, each with probability p_k the misclassification rate is:

$$\sum_{k=1}^{K} p_k(1 - p_k)$$

Does this agree with your expression above?