# Machine learning and statistical learning

**Support vector machines**

E. Gallic

## 1 Context

In this exercise, you will generate a binary response variable, and then train some SVM on that data. You will vary the values of the parameters in the estimations and look at the effect of the predicted classes.

## 2 Generate some data

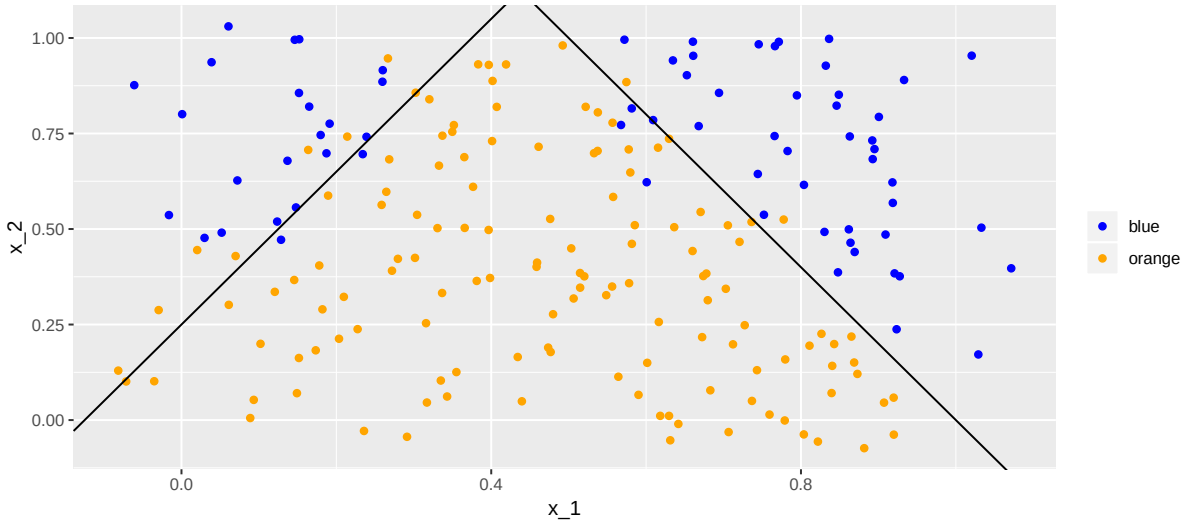


Figure 1: Example of generated points.

1. Generate 200 points from the unit square:
   - for example, draw the $x_1$ coordinates from a $\mathcal{U}[0,1]$ distribution and the $x_2$ coordinates from a $\mathcal{U}[0,1]$ .
2. Assign the label "blue" and "orange" to each of the 200 observations according to the following rule:
   - if $2 \times x_1 + 0.25 - x_2 > 0$ and $-2 \times x_1 + 2 - x_2 > 0$, then the label should be "orange"
   - otherwise, the label should be "blue".

3. Now, add a noise to each point. The noise should be randomly drawn from a $\mathcal{U}[-.1, .1]$ (this way, the 200 generated points should not be perfectly separable).

4. Graph your points on a scatter plot, where the color of each point should reflect its label. Add the true boundaries. The equations of the separating lines are : $x_2 = 2x_1 + 0.25$ and $x_2 = -2x + 2$.

## 3 Training the SVM algorithm

1. Using either `R` or Python, train a SVM classifier on your 200 points, picking a **linear kernel**. Then, visualize the resulting boundaries. What can you see?

2. Now, fit a SVM classifier using a **polynomial kernel**. According to the true form of your boundary, what value of the degree should you pick?

3. Using your trained classifier, predict the values on a grid ranging from $-.1$ to 1.1 for each dimension (for $x_1$ and for $x_2$).

4. Plot the dots of your grid on a scatter plot, and add the true boundaries.

5. Try different values for the cost parameter. For each cost parameter that you set, predict the values at each point of your grid and plot the points matching the color to the predicted class.

6. Assume that you do not know the true boundary. If you are to fit a SVM using a polynomial kernel, how would you do to pick a value for the cost parameter and for the degree? Explain you methodology in details.