Machine learning algorithms: exercise 1 16.03.2023

1. Data.txt file contains 212 two dimensional points. Points from location 1 to location 100 belong to class C_1, points from location 101 to location 200 belong to class C_2. The origin of the remaining 12 points is unknown. They belong to either class C_1 or C_2. Points from class C_1 and C_2 are shown in fig.1. Find the unit weight vector **w** that is perpendicular to line **l** that passes through points $\mathbf{p}_1$=(-2,6) and $\mathbf{p}_2$=(6,-2).

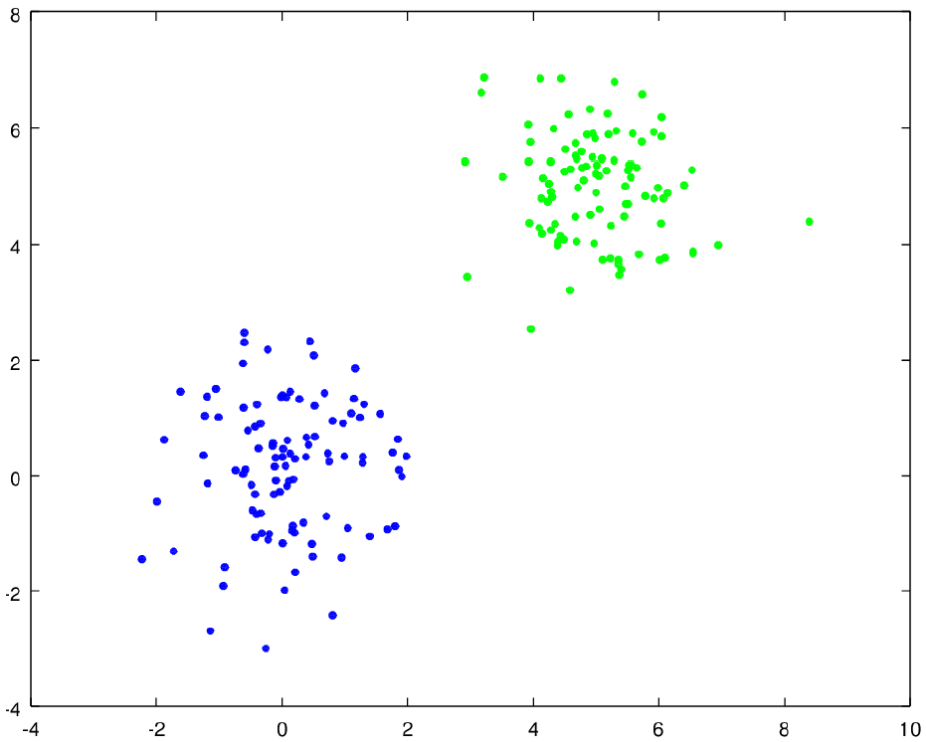

Figure 1. Points from class C_1 are blue and points from class C_2 are green.

2. Find such a threshold t for which $\mathbf{w}^T\mathbf{x}<t$ when **x** belongs to class C_1 and $\mathbf{w}^T\mathbf{x}>t$ when **x** belongs to class C_2. Use your threshold and classify remaining 12 points in file Data.txt to class C_1 and C_2. (Note $\mathbf{w}^T\mathbf{x}=\mathbf{w}\bullet\mathbf{x}$, where T is transpose operator).

3. Modify your classifier such that the classification can be done comparing the result to zero instead to that of t in previous task.

4. Calculate the projections of points from class C_1 and C_2 on the directions of **w** and **l**. Draw histograms for both directions and interpret your results.

5. Let us consider the points from class C_1 alone. (from position 1 to position 100). The probability to an event that a point $\mathbf{p}_1$ is within some range from point $\mathbf{p}_2$ can be considered as a function of distance $d(\mathbf{p}_1,\mathbf{p}_2)$ between the points. What is the probability for the event that a point in class C_1 belong to the circular area with center point of (0,0) and the radius that is the mean of all distances of points in C_1 from the center point (0,0)?

6. Fit a linear regression model to the data using points from 1 to 200. Inspect your result visually and consider what kind of problems you may encounter later if you use your model with new data.