Lecture 20 Part 2 - Soft-Margin SVM

Soft-Margin Support Vector Machine (SVM): Overlapping Classes

To handle this case, the SVM implementation has a bit of a fudge-factor which "softens" the margin: that is, it allows some of the points to creep into the margin if that allows a better fit. The hardness of the margin is controlled by a tuning parameter, most often known as **slack varible** $\xi_n \geq 0, n = 1, \ldots, N$, with one slack variable for each training data point. For very large ξ, the margin is hard, and points cannot lie in it. For smaller ξ, the margin is softer, and can grow to encompass some points.

A **slack variable** is defined as $\xi_n = 0$ for data points that are on or inside the correct margin boundary and $\xi_n = |t_n - y(x_n)|$ for other points. Thus a data point that is on the decision boundary $y(x_n) = 0$ will have $\xi_n = 1$, and points with $\xi_n > 1$ will be misclassified. The exact classification constraints are then replaced with

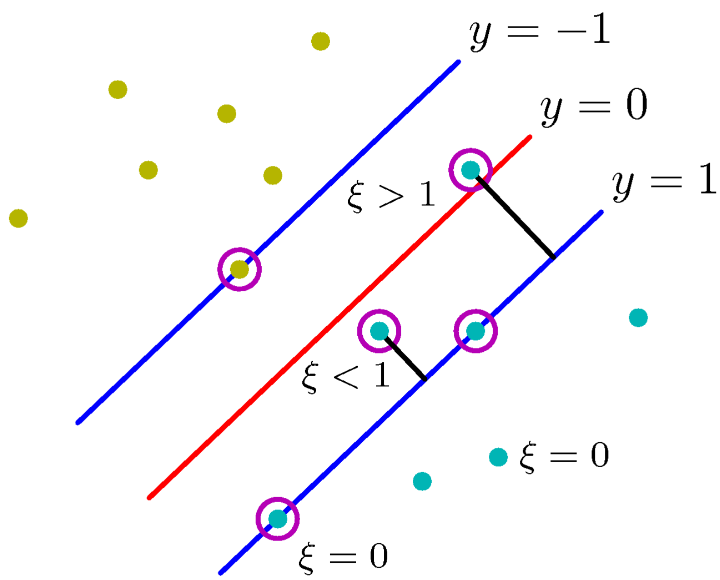
$$t_n y(x_n) \geq 1 - \xi_n, n = 1, \ldots, N$$

in which the slack variables are constrained to satisfy $\xi_n \geq 0$.

- Data points for which $\xi_n = 0$ are correctly classified and are either on the margin or on the correct side of the margin.
- Points for which $0 < \xi_n \leq 1$ lie inside the margin, but on the correct side of the decision boundary.
- And those data points for which $\xi_n > 1$ lie on the wrong side of the decision boundary and are misclassified.

In [1]:
```
from IPython.display import Image
Image('figures/Figure7.3.png', width=400)
```

Out[1]:



Our goal is now to maximize the margin while softly penalizing points that lie on the wrong side of the margin boundary. We therefore minimize:

$$\begin{gathered}\arg_{w,b}\min \frac{1}{2}\|w\|^2 + C\sum_{n=1}^{N}\xi_n \\ \text{subject to } t_n y(x_n) \geq 1 - \xi_n, n = 1, \dots, N \\ \text{and } \xi_n \geq 0, n = 1, \dots, N\end{gathered}$$

where the parameter $C > 0$ controls the trade-off between the slack variable penalty and the margin.

- Because any point that is misclassified has $\xi_n > 1$, it follows that $\sum_n \xi_n$ is an upper bound on the number of misclassified points.
- The parameter C is therefore analogous to (the inverse of) a regularization coefficient because it controls the trade-off between minimizing training errors and controlling model complexity.
- In the limit $C \to \infty$, we will recover the earlier support vector machine for separable data.

to be continued...
