MA2823: Foundations of Machine Learning **Homework 9**

Due December 2, 2016.

Question 1

Let us consider the training data $\{(\boldsymbol{x}^1, y^1), \ldots, (\boldsymbol{x}^n, y^n)\}$ where $\boldsymbol{x}^i \in \mathbb{R}^p$ and $y^i \in \{-1, +1\}$. A soft-margin SVM solves

$$\begin{aligned} \arg \min_{\boldsymbol{w}\in\mathbb{R}^p, b\in\mathbb{R}, \boldsymbol{\xi}\in\mathbb{R}^n} \quad & \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1}^{n} \xi_i \\ \text{s. t. } & y^i(\langle \boldsymbol{w}, \boldsymbol{x}^i \rangle + b) \geq 1 - \xi_i \\ & \xi_i \geq 0 \qquad\qquad \forall i \in \{1, \ldots, n\} \end{aligned} \tag{1}$$

(a) Is a soft-margin SVM more likely to overfit if C is large or small?

> **Solution:** If C is large, more importance is given to the error on the training set, and the SVM is more likely to overfit.

(b) Give one way of choosing C in practice.

> **Solution:** By cross-validation.

(c) What does this mean for a feature j if the solution w_j is close to 0?

> **Solution:** That this feature is uninformative. The class won't depend on this feature. (Note: we're talking about a feature weight w_j, not a (support) vector coefficient α_i.)

(d) Give an interpretation of the two terms of Equation (1): $||\boldsymbol{w}||^2$ and $\sum_{i=1}^{n} \xi_i$.

> **Solution:** $||\boldsymbol{w}||$ is the inverse of the margin.
> $\sum_{i=1}^{n} \xi_i$ is the sum of slacks ξ_i, which quantify the error for each misclassified training point.