Machine learning@Data science, 2017–Exercises.

Set 3. Due March 13, 2017

Problem 13 Let $(x_1, y_1), \ldots, (x_n, y_n)$ be data in $\mathbb{R}^d \times \{-1, 1\}$. Suppose that the data are *linearly separable*, that is, there exists a $w \in \mathbb{R}^d$ such that $y_i w^T x_i > 0$ for all $i = 1, \ldots, n$. The *margin* of such a vector is

$$\gamma(w) = \min_{i=1,\ldots,n} \frac{y_i w^T x_i}{\|w\|} .$$

Formulate a *convex optimization problem* whose solution is a vector w^* that classifies the data correctly (i.e., $y_i {w^*}^T x_i > 0$ for all $i = 1, \ldots, n$) and maximizes the margin. Show that the optimal solution w^* lies in the vector space spanned by the examples x_i for which the margin $\frac{y_i {w^*}^T x_i}{\|w^*\|}$ is minimal among all examples. (These are the so-called support vectors.)

Problem 14 Let $\mathcal{H}$ be the Hilbert space of all sequences $s = \{s_n\}_{n=0}^{\infty}$ satisfying $\sum_{n=0}^{\infty} s_n^2 < \infty$ with inner product $\langle s, t \rangle = \sum_{n=0}^{\infty} s_n t_n$. Consider the feature map $\Phi : \mathbb{R} \to \mathcal{H}$ that assigns, to each real number x, the sequence $\Phi(x)$ whose n-th element equals

$$(\Phi(x))_n = \frac{1}{\sqrt{n!}} x^n e^{-x^2/2} , \quad n = 0, 1, 2, \ldots .$$

Determine the kernel function $K(x, y) = \langle \Phi(x), \Phi(y) \rangle$ for $x, y \in \mathbb{R}$. (You may use the fact that $\sum_{n=0}^{\infty} x^n/n! = e^x$.)

Can you generalize the kernel so that it is defined on $\mathbb{R}^d \times \mathbb{R}^d$ instead of $\mathbb{R} \times \mathbb{R}$? What is the corresponding feature map?

Problem 15 Let $K_1, K_2 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be kernels. Prove that $K_1 + K_2$ and $K_1 K_2$ are also kernels.

Problem 16 Write a program that generates n independent pairs of random variables (X_i, Y_i) such that $\mathbf{P}\{Y_i = 0\} = \mathbf{P}\{Y_i = 1\} = 1/2$ and, conditionally on $Y_i = 0$, X is multivariate normal with mean $(0, 0, \ldots, 0)$ and identity covariance matrix, while, conditionally on $Y_i = 1$, X is multivariate normal with mean $(1, 1, 0, 0, \ldots, 0)$ and identity covariance matrix.

Train a decision-tree classifier that greadily splits each cell by minimizing the number of misclassified points until it has k cells and assigns a majority vote to each cell.

- Test the performance of the classifier on independent test data for a wide range of the parameters n, d, and k.

- Implement bagging for the decision-tree classifier above (by training the classifier of many subsamples and taking a majority vote) and, again, test its performance for a wide range of the parameters n, d, and k.

- Implement the random-subspace method that chooses two of the d components at random, builds the decision-tree classifier above, repeats this many times and takes a majority vote of the obtained classifiers. Test the performance for a wide range of the parameters n, d, and k.