Elements of Bioinformatics, Autumn 2011, Exercise set 4

Return your answers to Esa Pitkänen (firstname.lastname@cs.helsinki.fi) on writing before the beginning of the review session, Mon 28.11 at 10am at the latest.

1. Consider the following two data points $\mathbf{x}_1 = (-1, -1), y_1 = -1, x_2 = (1, 1), y_2 = 1$, and the following hyperplanes $\mathbf{w}_1 = (2, 2), b_1 = -1$ and $\mathbf{w}_2 = (1, 1), b_2 = 0$). $\mathbf{w}_3 = (-1/\sqrt{2}, -1/\sqrt{(2)}), b_3 = 1/\sqrt{2}$

   Compute the following for all hyperplanes and both data points

   (a) Functional margin $m(\mathbf{x}) = y(\langle \mathbf{w}, \mathbf{x} \rangle + b)$

   (b) Geometric margin $\gamma(\mathbf{x}) = y(\langle \frac{\mathbf{w}}{||\mathbf{w}||}, \mathbf{x} \rangle + \frac{b}{||w||})$

   Identify the maximum margin hyperplane. Which of the above notions of the margin is the useful one?

2. On slide 18 of Lecture 5, it is shown how plugging in the equation $\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$ for the weight vector makes the optimization problem only depend on the inner products of the feature vectors. Derive the following equations used on the slides (show the intermediate steps):

   (a) $||\mathbf{w}||^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle$

   (b) $\langle \mathbf{w}, \mathbf{x}_i \rangle = \sum_{j=1}^{n} \alpha_i y_i \langle \mathbf{x}_i, \mathbf{x}_j \rangle$

3. Consider the following set of data points $\mathbf{x}_1 = (-1, -1), y_1 = 1, \mathbf{x}_2 = (+1, +1), y_2 = 1, \mathbf{x}_3 = (-1, +1), y_3 = -1, \mathbf{x}_4 = (+1, -1), y_4 = -1$

   (a) What is the minimum number of misclassified examples that can be achieved by a hyperplane?

   (b) Devise a mapping of the data into another feature space where the data becomes linearly separable.

4. Consider a dataset of 100 training examples with features corresponding to expression levels of 6000 genes. Suppose that we wish to train an SVM using

   - Linear kernel $k_1(\mathbf{x}, \mathbf{x}') = \langle \mathbf{x}, \mathbf{x}' \rangle$
   - Quadratic kernel $k_2(\mathbf{x}, \mathbf{x}') = (\langle \mathbf{x}, \mathbf{x}' \rangle + 1)^2$
   - Cubic kernel $k_3(\mathbf{x}, \mathbf{x}') = (\langle \mathbf{x}, \mathbf{x}' \rangle + 1)^3$

   What are the dimensions of the feature vectors represented by the kernels? How much time would computing the kernel matrix (all-pairwise kernel values $k(x_i, x_j)$) take if the kernels were to computed by first writing down the feature vectors and computing the inner product of the feature vectors, compared to using the above equations.

5. There are many examples commonly of similarity scores used in bioinformatics that are not valid kernels (e.g BLAST bit-score) meaning that they cannot be expressed in terms of inner products of feature vectors. If such scores are used in place of valid kernels, there might be several local optima which the SVM optimizer could fall instead of finding the globally optimal solution, resulting in a model that has inferior predictive accuracy.

   However, given a set of examples $x_1, \ldots, x_n$) a non-kernel similarity score can be converted into a feature vector and subsequently a kernel by

$$\phi(x) = (S(x, x_1), S(x, x_2), \ldots, S(x, x_n)),$$

   where $S(x, x')$ is the similarity score between x and x'.

   Discuss the pros and cons of using this approach with SVMs. How does the training set affect the kernel?