CS 680

Homework 2 : SVMs

(due October 4th)

1. **SVM with no bias term [35 pts].**
   Formulate a soft-margin SVM without the bias term, i.e. $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$. Derive the saddle point conditions, KKT conditions and the dual. Compare it to the standard SVM formulation. What is the implication of the difference on the design of SMO-like algorithms? Recall that SMO algorithms work by iteratively optimizing two variables at a time. Hint: consider the difference in the constraints.

   Discuss the merit of the bias-less formulation as the dimensionality of the data (or the feature space) is varied. When using this SVM formulation it may be useful to add a constant to the kernel matrix. Explain why this can be beneficial.

2. **Soft-margin for separable data [15 pts].**
   Consider training a soft-margin SVM with C set to some positive constant. Suppose the training data is linearly separable. Since increasing the ξ_i can only increase the objective of the primal problem (which we are trying to minimize), at the optimal solution to the primal problem, all the training examples will have functional margin at least 1 and all the ξ_i will be equal to zero. True or false? Explain! Given a linearly separable dataset, is it necessarily better to use a a hard margin SVM over a soft-margin SVM?

3. **In-bound SVs in soft-margin SVMs [20 pts].**
   Examples $\mathbf{x}_i$ with $\alpha_i > 0$ are called *support vectors* (SVs). For soft-margin SVM we distinguish between *in bound* SVs, for which $0 < \alpha_i < C$, and *bound* SVs for which $\alpha_i = C$. Show that in-bound SVs lie exactly on the margin. Argue that bound SVs can lie both on or in the margin, and that they will "usuallly" lie in the margin. Hint: use the KKT conditions.

4. **Margin of optimal margin hyperplanes [10 pts].**
   In class we saw that the geometric margin of an optimal margin hyperplane, denoted by ρ, is equal to $2/||\mathbf{w}||$. Show that for an optimal margin hyperplane

   $$2\rho^{-2} = \sum_{i=1}^{n} \alpha_i$$

   and

   $$4\rho^{-2} = 2W(\alpha) = ||\mathbf{w}||^2 ,$$

   where $W(\alpha)$ is the dual function to be optimized.

5. **Some experiments [20 pts].**
   In this question we will explore some practical aspects of SVM use. Download the `heart` dataset from the homework page.

   - Obtain results of cross-validation on the dataset using an SVM with a linear kernel and default setting of the C parameter. Note that the `log` attribute of the `Results` provides information on the training procedure: how long it took, how many support vectors resulted etc. Repeat the experiment after *standardizing* the data, i.e. rescaling each variable to have 0 mean and unit variance using the `preproc.Rescale` object (see the PyML tutorial for explanation on how to use it). What are the differences you observe between the training on the normalized vs. non-normalized data? Observe what happens as you vary the soft margin parameter. Repeat this experiment using the Gaussian kernel.
   - Another common form of normalization is to project the data to the unit sphere, i.e. normalize each example to be a unit vector by dividing each component by the norm of the vector. This is implemented in the context of a kernel by using a normalized kernel of the form $K(\mathbf{x}, \mathbf{x}')/\sqrt{K(\mathbf{x}, \mathbf{x})K(\mathbf{x}', \mathbf{x}')}$. Explain why using this cosine-like kernel is not necessary when using a Gaussian kernel.

6. **Understanding the effect of SVM parameters [15 pts].**
   Use the `demo2d` module of `PyML` to construct a two-dimensional dataset. Create a series of plots that illustrate the points made in class regarding the effects of SVM parameters on the resulting decision surface. Submit the data you created, as well as the code snippets used to generate the figures, so that your work will be reproducible: The best series of experiments will be used in the SVM-howto.