IEOR E4525
Martin Haugh

Due: Monday 7^{th} March 2016

Assignment 4

1. *(Scaling the Inputs)*
   True or false: in training an SVM it is generally a good idea to scale all input variables so that, for example, they all lie in some fixed interval or so that they all have the same mean, μ, and variance, σ^2, e.g $(\mu, \sigma^2) = (0, 1)$. Justify your answer.

2. *LIBSVM* is an integrated software library for SVM classification, regression and distribution estimation / novelty detection. It is available in *R*, *Matlab*, *Python* etc. and a detailed description of the library can be found at

   `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`.

   It can be installed in *R* by simply installing the *e1071 package*. Installing it in *Matlab* requires more work (and can be troublesome) but see the above url for details.

   Familiarize yourself with the use of *LIBSVM* for binary classification, multi-class classification and regression. Note also how to use it for for parameter selection via cross-validation. How does it handle categorical data? How does it handle missing data? If most of the data-points had at least one missing component, how might you address this? Does it scale the data?

3. *(Classifying Tumors)*

   (a) Install the breast-cancer data that is part of the *mlbench* (machine-learning benchmark problems) package in *R*. Use the functionality of *SVMLIB* to build an SVM classifier for this data. You should randomly assign $t\%$ of your data to the training set and the remainder of your data to the test set. Then use cross-validation on your training set to build your classifier. You can take $t = 70\%$ initially.

   (b) Repeat part (a) $N = 50$ times to get N samples of the performance of the trained classifier on the test set. (Note that each of the N samples will have different training and test sets.) Compute the mean and standard deviation of the test-set-performance.

   (c) Repeat part (b) for values of $t = 50\%, 55\%, \ldots, ...95\%$ and plot the mean test-set performance together with 95% confidence intervals for this performance against t. What conclusions can you draw?

4. *(SVMs and Cross-Validation)*
Suppose you have successfully trained an SVM with 10,000 training points and a Gaussian kernel where the values of C and σ were selected via cross-validation. Recall that the Gaussian kernel has the form

$$k(\mathbf{x}, \mathbf{x}') \;=\; \exp\left(-\frac{||\mathbf{x}-\mathbf{x}'||^2}{2\sigma^2}\right).$$

You are then given an additional 40,000 training points and so you wish to retrain your SVM using the entire 50,000 training points that you now have. However, you wish to avoid the heavy computational expense associated with repeating the cross-validation exercise and so you therefore simply choose the SVM using the values of C and σ that you obtained with the earlier 10,000 training points. Do you see any potentially major problem with this? If so, what is it?

5. *(Support Vector Regression (SVR))*
We claimed in class that the SVR problem could be reduced to solving the following **primal** problem formulation:

$$\begin{aligned}
\min_{\mathbf{w},\, b,\, \boldsymbol{\xi}\geq 0,\, \hat{\boldsymbol{\xi}}\geq 0} \quad & C\sum_{i=1}^{n}\left(\xi_i + \hat{\xi}_i\right) \;+\; \frac{1}{2}||\mathbf{w}||^2 \\
\text{subject to} \quad t_i \;&\leq\; y(\mathbf{x}_i) + \epsilon + \xi_i\,, \qquad i = 1,\ldots,n \qquad (1)\\
t_i \;&\geq\; y(\mathbf{x}_i) - \epsilon - \hat{\xi}_i\,, \qquad i = 1,\ldots,n. \qquad (2)
\end{aligned}$$

(a) Justify this formulation.

(b) Introduce Lagrange multipliers $a_i \geq 0$ and $\hat{a}_i \geq 0$ for (1) and (2), respectively, and $\mu_i \geq 0$ and $\hat{\mu}_i \geq 0$ for the constraints $\xi_i \geq 0$ and $\hat{\xi}_i \geq 0$, respectively. Now optimize the Lagrangian with respect to $\mathbf{w}$, b, $\boldsymbol{\xi}$ and $\hat{\boldsymbol{\xi}}$.

(c) Let $L(\mathbf{a}, \hat{\mathbf{a}}, \boldsymbol{\mu}, \hat{\boldsymbol{\mu}})$ be the optimized Lagrangian. Now explain why the solution to

$$\max_{\mathbf{a}\geq 0,\, \hat{\mathbf{a}}\geq 0,\, \boldsymbol{\mu}\geq 0,\, \hat{\boldsymbol{\mu}}\geq 0} \; L(\mathbf{a}, \hat{\mathbf{a}}, \boldsymbol{\mu}, \hat{\boldsymbol{\mu}}) \qquad (3)$$

provides a lower bound on the optimal objective function of the primal problem. The problem in (3) is in the fact the **dual** problem. (And because the primal problem is a "well-behaved" convex quadratic problem we know the bound is tight so that the optimal value of the primal is equal to the optimal value of the dual problem.)

(d) Show that the optimal solution to the primal problem takes the form

$$y(\mathbf{x}) \;=\; \sum_{i=1}^{n}(a_i - \hat{a}_i)\mathbf{x}^\top\mathbf{x}_i + b.$$

(e) Use the KKT conditions to see that for every data-point $\mathbf{x}_i$, either $a_i = 0$ or $\hat{a}_i = 0$ or $a_i = \hat{a}_i = 0$. What are the data-points that contribute to the predicted target value of a new data-point, $\mathbf{x}$?

(f) Given an optimal solution to the dual problem, how would you calculate the optimal value of b?

6. *(Kernels and LDA: Exercise 12.10 from HTF)*
Suppose you wish to carry out a linear discriminant analysis (LDA) with two classes using a vector of transformations of the input variables, $\phi(\mathbf{x})$. Since $\phi(\mathbf{x})$ is high-dimensional, you will use a regularized within-class covariance matrix $\mathbf{W}_\phi + \gamma\mathbf{I}$. Show that the model can be estimated using only the inner products $K(\mathbf{x}_i, \mathbf{x}_j) =< \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) >$. Hence the kernel property of SVM is also shared by regularized LDA.

Hint: The *Woodbury* or *Sherman-Morrison* formula might help.