Homework 4: October 28, 2016

Machine Learning

Susan Cherry

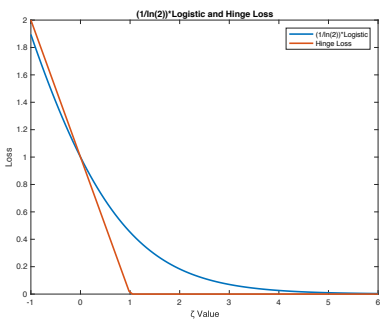
Question 1

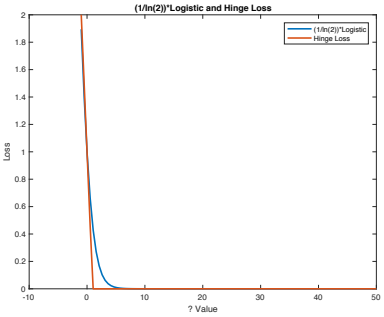
Part A: In this question, the kernel is just the simple linear function $k(x_i, x_j) = x_i^T x_j$, where $k(\cdot, z) : R^n \to R$ and $k(\cdot, z) = (\cdot^T z)$. Define $f(x) = \theta^T x$ and $g(x) = \lambda^T x$. Define the inner product $< f, g >_{H_k} = \theta^T \lambda$. Since this is just the simple linear kernel, we know from the SVM notes that it is symmetric, bilinear, and positive semi definite. Notice that $< f, f >_{H_k} = \theta^T \theta$. The norm is $||f||_{H_k} = (< f, f >)^{\frac{1}{2}}$, which is the norm required by the $||f||^2_{H_k} = \theta^T \theta$ in the problem.

From our notes, we know $H_k = \{f : f = \sum_i \alpha_i k(\cdot, x_i)\}$. In this case, we have $H_k = \{f : f = (\theta^T \cdot)\}$ where $f : R^n \to R$ and $k : R^n \to R$. Next I need to show the reproducing property, or that $< k(\cdot, x, f) >_{H_k} = \theta^T x = f(x)$ which is just the evaluation of f at x. Thus, k has the reproducing property. I'll also formally define the norm $||f||_{H_k} = (< f, f >_{H_k})^{\frac{1}{2}}$. So we have $H_k = \{f : f = (\cdot)^T \theta_i\}$. I've defined everything necessary to define the Reproducing Kernel Hilbert Space.

By the Representer Theorem, we know that the solutions of the optimization problem $f^* \in argmin_{f \in H_k} \sum_i \ell(f(x_i), y_i) + C||f||^2_{H_k}$ can be expressed as $f^*(x) = \sum_i \alpha_i k(\theta, x) = \sum_i \alpha_i (\theta^T x)$ or more generally, $f^*(\cdot) = \sum_i \alpha_i k(\theta, \cdot) = \sum_i \alpha_i (\theta^T \cdot)$. This shows that for the SVM optimization problem we only need to solve for α_i, which agrees with the solution from the Lagrangian for the SVM.

Part B: Below is a plot of the logistic and hinge loss. You can see that the hinge loss decreases linearly until it reaches 1. At this point it becomes and stays 0. Logistic loss also decreases asymptotically to 0, though not linearly. At ζ=5, logistic loss has decreased enough to be effectively 0. Thus, logistic loss decreases at a slower rate than hinge loss, because hinge loss reached 0 at ζ=1. The plot to the right is just a zoomed out version of the plot to the left. By zooming out, we can see the the hinge loss and logistic loss look very similar for most of the plot.





Part C: Now we have the primal problem which leads to the following Lagrangian: $L(\theta, \theta_0, \zeta, \alpha) = \frac{1}{2}||\theta||^2 + C\sum_{i=1}^{n} ln(1+exp(-\zeta_i)) + \sum_{i=1}^{n} \alpha_i(-y_i(\theta^T x_i + \theta_0) + \zeta_i)$.

Writing the KKT Conditions, I find the following:
$\nabla_\theta L(\theta, \theta_0, \zeta, \alpha) = \theta - \sum_i \alpha_i y_i x_i = 0$ which leads to $\theta = \sum_i \alpha_i y_i x_i$
$\nabla_{\theta_0} L(\theta, \theta_0, \zeta, \alpha) = -\sum_i \alpha_i y_i = 0$ which leads to $\sum_i \alpha_i y_i = 0$
$\nabla_{\zeta_i} L(\theta, \theta_0, \zeta, \alpha) = -C\frac{1}{1+e^{\zeta_i}} + \alpha_i = 0$. This leads to $\zeta_i = ln(\frac{C}{\alpha_i} - 1)$.
Dual Feasibility: $\alpha_i \geq 0$ for all i.
Complementary Slackness: $\alpha_i(-y_i(\theta^T x + \theta_0) + \zeta_i) = 0$ for all i.
Primal Feasibility $-y_i(\theta^T x + \theta_0) + \zeta_i \leq 0$

Using these conditions, $L(\theta, \theta_0, \zeta, \alpha) = \frac{1}{2}||\theta||^2 + C\sum_{i=1}^{n} ln(1 + exp(-\zeta_i)) - \sum_{i=1}^{n} \alpha_i(y_i(\theta^T x_i) - \sum_{i=1}^{n} \alpha_i(y_i\theta_0) + \sum_{i=1}^{n} \alpha_i\zeta_i$
$= \frac{1}{2}||\theta||^2 - ||\theta||^2 + C\sum_{i=1}^{n} ln(1 + exp(-\zeta_i)) + \sum_{i=1}^{n} \alpha_i\zeta_i$
$= -\frac{1}{2}\sum_i \sum_j \alpha_i\alpha_j y_i y_k x_i^T x_k + \sum_i \alpha_i(\frac{C-\alpha_i}{\alpha_i}) + C\sum_i ln(\frac{C}{C-\alpha_i})$.

So the dual problem is $max_\alpha -\frac{1}{2}\sum_i \sum_j \alpha_i\alpha_j y_i y_k x_i^T x_k + \sum_i \alpha_i(\frac{C-\alpha_i}{\alpha_i}) + C\sum_i ln(\frac{C}{C-\alpha_i})$.
Note that since $\zeta_i = ln(\frac{C}{\alpha_i} - 1)$, $\frac{C}{\alpha_i} - 1 \geq 0$ implying that $\alpha_i < C$ and $0 < \alpha_i < C$.
Thus we have the constraints $0 < \alpha_i < C$ for all i and $\sum_i y_i\alpha_i = 0$. The $0 < \alpha_i < C$ is an interesting result.

Compare it to the SVM constraints of $0 \leq \alpha_i \leq C$. In the logistic regression case, the inequalities are strict, meaning that α_i is always greater than 0. For SVM, though, actually most of the α_i's will be zero meaning that only the support vectors will determine the function. For logistic regression, all α_i determine the function meaning that logistic regression is sensitive to outliers.

Furthermore, we can rearrange and rewrite the $\sum_i \alpha_i(\frac{C-\alpha_i}{\alpha_i}) + C\sum_i ln(\frac{C}{C-\alpha_i}) = -C\sum_i[ln(1-\frac{\alpha_i}{C}) + \frac{\alpha_i}{C}ln(\frac{\alpha_i}{C}) - \frac{\alpha_i}{C}ln(\frac{C-\alpha_i}{C})] = -C\sum_i[\frac{\alpha_i}{C}ln([\frac{\alpha_i}{C}) + (1-\frac{\alpha_i}{C})ln((1-\frac{\alpha_i}{C})] = C\sum_i H[\frac{\alpha_i}{C}, \frac{1-\alpha_i}{C}]$, which is the Bernoulli Entropy. This is different than the SVM case in which we had $C\sum_i \frac{\alpha_i}{C}$ which is like a probability, not entropy.

Question 2

Part A: Suppose we have two data points. One is labeled $y_1 = 1$ and and one is labeled $y_2 = -1$. We also have the feature vectors x_1 and x_2. Because these are the only two data points, we know that they are linearly separable. Also, since we are maximizing the minimum margin and there are only two data points, the distance from each point to the margin must be equal. If the distance for one was smaller than the other, the margin could be shifted away until the distance from both points was equal in order to maximize the minimum margin. This means that both points are support vectors and gives us the constraints. Let's use the scaled margin. We have the following Lagrangian: $L(\lambda, \lambda_0, \alpha) = \frac{1}{2}\sum_{i=1}^{2} \lambda_i^2 + \sum_i \alpha_i(y_i(\lambda^T x_i + \lambda_0) + 1)$.
Taking the KKT Conditions, we get:
1) $\nabla_\lambda L(\lambda, \lambda_0, \alpha) = \lambda - \alpha_1 y_1 x_1 - \alpha_2 y_2 x_2 = 0$, which leads to $\lambda = \alpha_1 x_1 - \alpha_2 x_2$.
2) $\nabla_{\lambda+0} L(\lambda, \lambda_0, \alpha) = \alpha_1 - \alpha_2 = 0$ which leads to $\alpha_1 = \alpha_2$.
3) Dual Feasibility: $W\alpha_i > 0$ for all $i = 1 and 2$.
4) Complementary Slackness: $\alpha_i(y_i(\lambda^T x_i + \lambda_0) + 1) = 0$ for $i = 1$ and 2.
5) Primal Feasibility :$\alpha_i(y_i(\lambda^T x_i + \lambda_0) + 1) \leq 0$ for $i = 1$ and 2.

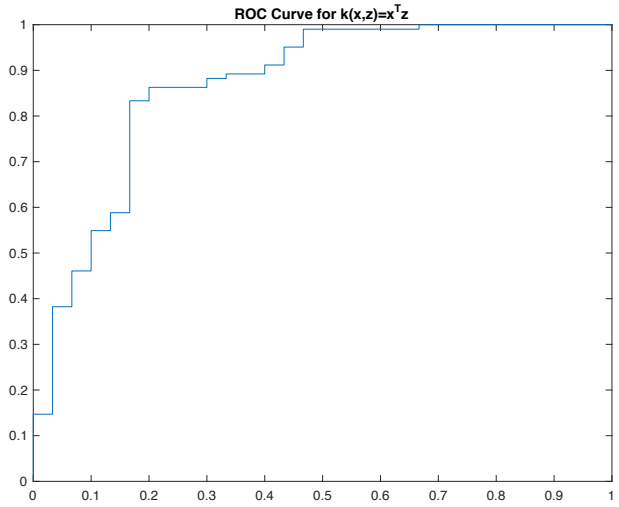
Now we know that $-y_1(\lambda^{*T} x_1 + \lambda_0^*) + 1{=}0$ and $y_2(\lambda^{*T} x_2 + \lambda_0^*) + 1 = 0$. Thus, $(\lambda^{*T} x_1 + \lambda_0^*){=}1$ and $-(\lambda^{*T} x_2 + \lambda_0^*) = 1$. Next I'll add these two and get $\lambda^T(x_2 - x_1) = -2$. Recall that $\alpha_1 = \alpha_2$ and $\lambda^* = \alpha_1^* y_1 x_1 + \alpha_2^* y_2 x_2 = \alpha_1^* x_1 - \alpha_2^* x_2 = \alpha_1^*(x_1 - x_2)$. Substituting, we get $\alpha_1(x_1 - x_2)^T(x_1 - x_2) = 2$ which leads to $\alpha_1 = \alpha_2 = \frac{2}{(x_1 - x_2)^T(x_1 - x_2)}$. This gives us the values $\lambda^* = \frac{2(x_1 - x_2)}{(x_1 - x_2)^T(x_1 - x_2)}$ and $\lambda_0^* = 1 - \frac{2(x_1 - x_2)^T}{(x_1 - x_2)^T(x_1 - x_2)} * x_1$ We can see from this problem that just two points (one from each class) are sufficient for determining the maximum margin hyperplane.

Part B: I need to show that finding the maximum margin hyperplane is a convex optimization problem. To do that, I need to show that the objective function and all of the constraints are convex. First, consider the objective function, which is $\frac{1}{2}||\lambda||_2^2 = \frac{1}{2}\sum_i \lambda_i^2$. The second derivative of $\frac{1}{2}\sum_i \lambda_i^2 = 1 > 0$ so this is convex. Thus objective function is the sum of convex functions meaning that it itself is convex Our constraints are $-y_i(\lambda^T x_i + \lambda_0) + 1 \leq 0$. For each i, this constraint is affine. We know that affine functions are convex. We also know that the α_i values are nonnegative. Therefore the constraints are all convex. So the we simply have the Lagrangian as the sum of convex functions, which is also convex. The maximum of a set of convex functions is again convex. So finding the maximum margin hyperplane is actually a convex optimization problem.

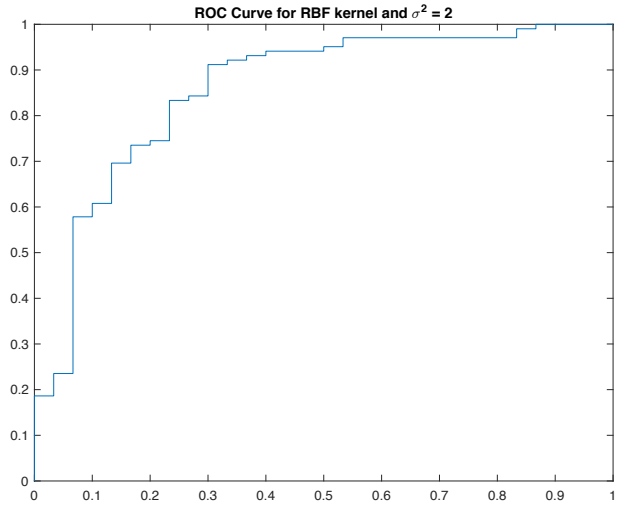
Question 3

Part A: See my uploaded code for the train and predict functions. I used the quadprog function in Matlab and tested using the first 100 observations of the Iris dataset. Everything seems to be working correctly.

Part B: Next I use the credit card data set and implement a soft margin kernel with $\mathrm{k}(x, z) = x^T z$. See my Matlab code for implementation. I randomly shuffled the data and trained the algorithm on 9/10 of the data and tested on the remaining 1/10. I found that the AUC is 0.8673. The misclassification error is 11.97%. Below is the ROC curve.



Part C: Next I use the radical basis kernel with $\sigma^2 = 2$. Note that I standardized my data. The AUC is now 0.8575 and the misclassification rate is 18.72%. The ROC curve is below.



Then I change to $\sigma^2 = 20$. Now the AUC is 0.8974, which is the highest of the three models and the misclassification rate is 11.98% (nearly identical to the

misclassification error in Part B). The ROC curve is below. Note that these results are probably highly dependent on how my data was randomly shuffled.