Duke University **November 2, 2016**

CS 571D

HW 4

1. • Elements in our RKHS must be in the span of the $f(x_i)$'s. Thus, we start out with our RKHS equal to $H = \{f(x) = \boldsymbol{\theta}^T x \mid x \in \mathcal{X}\}$.

   Now, we define a kernel in H as follows: $k(x, z) = x^T z$.

   We see that this is just an inner product, so it clearly satisfies all the properties of inner products.

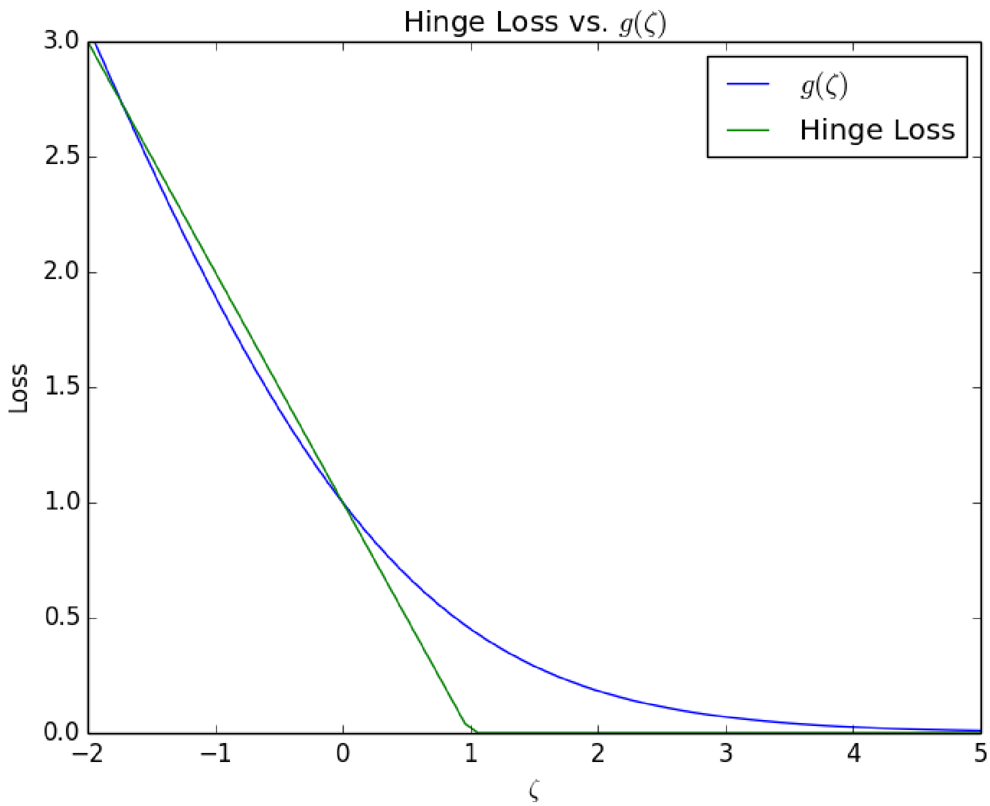
   Now, let's show that k is a reproducing kernel:

$$\langle k(., x), f \rangle = \boldsymbol{\theta}^T x = f(x)$$

   Since we have defined H and defined a reproducing kernel on H, we conclude that H is a RKHS.

   By the representer theorem, we know the optimal solution is in the form

$$f^* = \sum_{i=1}^{n} \alpha_i x_i^T \cdot$$



- b
- c: The ℓ_2 regularized logistic regression problem can be written as the following primal problem:

$$\min_{\boldsymbol{\theta},\theta_0,\zeta} \frac{1}{2}||\boldsymbol{\theta}||^2 + C\sum_{i=1}^{n} g(\zeta_i)$$

subject to:

$$y_i(f(x_i) + \theta_0) \geq \zeta_i, \forall i$$

This is equivalent to

$$\min_{\boldsymbol{\theta},\theta_0,\zeta} \frac{1}{2}||\boldsymbol{\theta}||^2 + C\sum_{i=1}^{n} g(\zeta_i) \;\text{ s.t. }\; -y_i((f(x_i) + \theta_0) - \zeta_i) \leq 0 \forall i$$

This leads to the Lagrangian

$$\mathcal{L}([\boldsymbol{\theta}, \theta_0], \boldsymbol{\alpha}) = \frac{1}{2}\sum_{j=1}^{n} \theta_j^2 + C\sum_{i=1}^{n} g(\zeta_i) + \sum_{i=1}^{n} \alpha_i \left[-y_i(f(x_i) + \theta_0) - \zeta_i\right]$$

We can write the KKT conditions, starting with Lagrangian stationarity.

$$
\begin{aligned}
&\nabla_{\boldsymbol{\theta}}\mathcal{L}([\boldsymbol{\theta},\theta_0],\boldsymbol{\alpha}) = \boldsymbol{\theta} - \sum_{i=1}^{n}\alpha_i y_i \mathbf{x}_i = 0 \implies \boldsymbol{\theta} = \sum_{i=1}^{n}\alpha_i y_i \mathbf{x}_i \\
&\frac{d}{d\lambda_0}\mathcal{L}([\boldsymbol{\theta},\theta_0],\boldsymbol{\alpha}) = \sum_{i=1}^{n}\alpha_i y_i = 0 \implies \sum_{i=1}^{n}\alpha_i y_i = 0 \\
&\forall i,\ \frac{d}{d\zeta_i}\mathcal{L} = \frac{Ce^{-\zeta_i}}{1+e^{-\zeta_i}} - \alpha_i = 0 \implies \zeta_i = \ln(\frac{c}{a} - 1) \\
&\alpha_i \geq 0 \ \ \forall i \ \ \text{(dual feasibility)} \\
&\alpha_i[-y_i(\boldsymbol{\theta}^T x_i + \theta_0) + \zeta_i] = 0 \ \ \forall i \ \ \text{(complementary slackness)} \\
&-y_i(\boldsymbol{\theta}^T x_i + \theta_0) + \zeta_i \leq 0 \ \ \forall i \ \ \text{(primal feasibility)}
\end{aligned}
$$

Using the KKT conditions, we can simplify the Lagrangian to get an expression for the dual:

$$
\begin{aligned}
\mathcal{L} &= \frac{1}{2}||\boldsymbol{\theta}||_2^2 + C\sum_{i=1}^{n} g(\zeta_i) + \boldsymbol{\theta}^T\sum_{i=1}^{n}(-\alpha_i y_i x_i) + \sum_{i=1}^{n}(-\alpha_i y_i \theta_0) + \sum_{i=1}^{n}\alpha_i\zeta_i \\
&= \frac{1}{2}||\boldsymbol{\theta}||^2 + C\sum_{i=1}^{n}\left[\ln\left(1+e^{-\zeta_i}\right)\right] + \sum_{i=1}^{n}\zeta_i\alpha_i \\
&= \frac{1}{2}||\boldsymbol{\theta}||^2 + C\sum_{i=1}^{n}\left[\ln\left(1+e^{-\ln\left(\frac{C}{\alpha_i}-1\right)}\right)\right] + \sum_{i=1}^{n}\alpha_i\ln\left(\frac{C}{\alpha_i}-1\right) \\
&= \frac{1}{2}||\boldsymbol{\theta}||^2 + C\sum_{i=1}^{n}\left[\ln\frac{C}{\alpha_i}\right] + \sum_{i=1}^{n}\alpha_i\ln\left(\frac{C}{\alpha_i}-1\right) \\
&= \frac{1}{2}||\boldsymbol{\theta}||^2 + nC\ln C + \sum_{i=1}^{n}(\alpha_i - C)\ln(C-\alpha_i) - \sum_{i=1}^{n}\alpha_i\ln\alpha_i
\end{aligned}
$$

We can now formulate the dual problem:

$$\max_{\boldsymbol{\alpha}}\mathcal{L}(\boldsymbol{\alpha})$$

where

$$
\begin{aligned}
&\mathcal{L}(\boldsymbol{\alpha}) = -\frac{1}{2}\sum_{i,k}\alpha_i\alpha_k y_i y_k x_i^T x_k + \sum_{i=1}^{n}(\alpha_i - C)\ln(C-\alpha_i) - \sum_{i=1}^{n}\alpha_i\ln\alpha_i \\
&\text{s.t. } \ \forall i\ \alpha_i \geq 0 \,\text{and}\, 0 \leq \alpha_i \leq C, \sum_{i=1}^{n}\alpha_i y_i = 0
\end{aligned}
$$

This is similar to the dual for the non-separable case of SVM in that the αs are bounded by C on top, but different in that there are more terms in the lagrangian. Specifically, there is an additional penalty for increasing the αs, and the sum over the αs has now been modified to involve a log and the C term.

2. a. Solving the optimization problem in SVM is equivalent to finding the maximum margin hyperplane, as shown in the notes. Therefore, if we can solve the SVM optimization problem for a data set of just two points, then two points are sufficient to find the maximum-margin hyperplane.

The SVM optimization problem with two points is the following:

$$\text{min}_{\boldsymbol{\lambda},\lambda_0}\frac{1}{2}||\boldsymbol{\lambda}||_2^2$$

subject to:

$$\boldsymbol{\lambda}^T x_1 + \lambda_0 + 1 \leq 0, -\boldsymbol{\lambda}^t x_2 - \lambda_0 + 1 \leq 0$$

From here, we can form the lagrangian:

$$\mathcal{L}([\boldsymbol{\lambda}, \lambda_0], \boldsymbol{\alpha}) = \frac{1}{2}\sum_{j=1}^{p}\lambda_j^2 + \alpha_1(\boldsymbol{\lambda}^t x_1 + \lambda_0 + 1) + \alpha_2(-\boldsymbol{\lambda}^T x_2 - \lambda_0 + 1)$$

We can take the gradient with respect to $\boldsymbol{\lambda}$:

$$\nabla_{\boldsymbol{\lambda}}\mathcal{L} = \boldsymbol{\lambda} + \alpha_1 x_1 - \alpha_2 x_2 = 0 \implies \boldsymbol{\lambda} = \alpha_2 x_2 - \alpha_2 x_1$$

We can take the derivative with respect to λ_0 and set it equal to zero, which implies that $\alpha_1 = \alpha_2$.

This also tells us that $\lambda_0 = 1 - \boldsymbol{\lambda} x_1$.

We can take the gradient with respect to $\boldsymbol{\alpha}$ and set it equal to zero, obtaining the equations $\alpha_1 = \alpha_2 = \frac{2}{||x_1 - x_2||_2^2}$

Now, we have solved the problem. We have:

$$\begin{aligned}
\boldsymbol{\lambda} &= \alpha_1 y_1 x_1 + \alpha_2 y_2 x_2 \\
&= \frac{2}{||x_2 - x_1||_2^2}(x_1 - x_1) \\
\lambda_0 &= \frac{||x_2||^2 - ||x_1||^2}{||x_1 - x_2||^2}
\end{aligned}$$

b. We call an optimization problem convex if it is in the form:

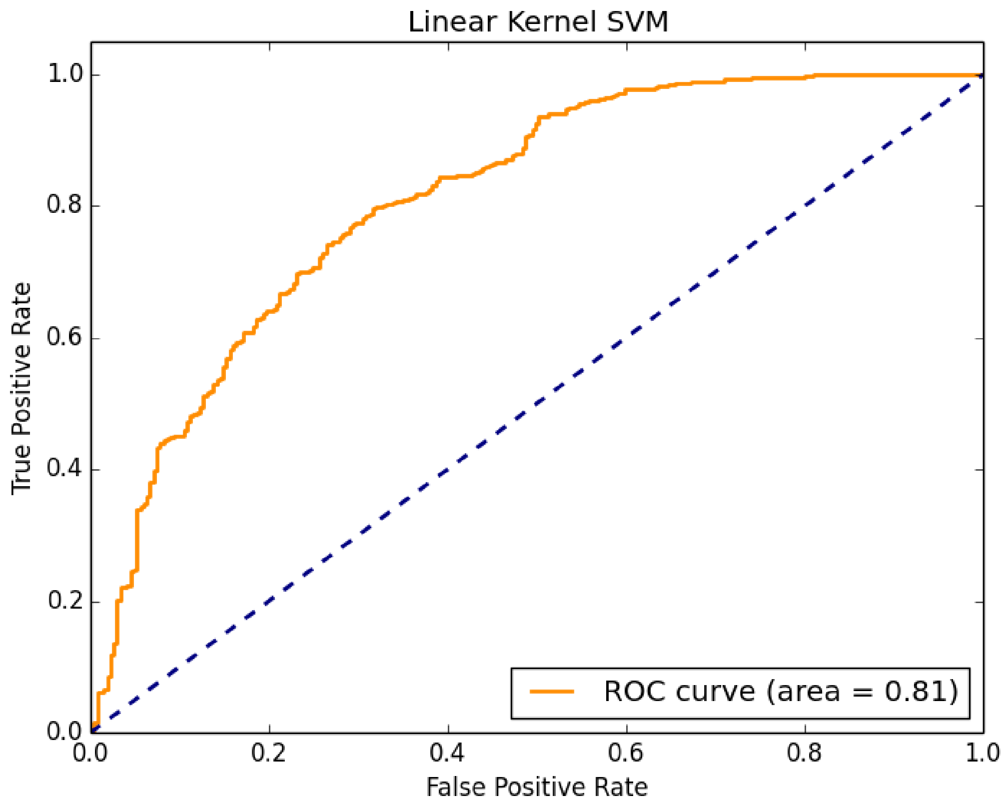
$$\begin{aligned} &\text{minimize} && f(x) \\ &\text{subject to} && g_i(x) \leq 0, i = 1, \ldots, n \end{aligned}$$

and the functions $f(x), g_1(x), \ldots, g_n(x)$ are all convex. In our case, $f = \frac{1}{2}||\lambda||_2^2$ and $g_i(x) = y_i(\lambda^T x_i = \lambda_0) + 1 \leq 0$. We have that a function is convex if the Hessian H of the function is positive semidefinite, where $H_{i,j} = \frac{d^2 f}{dx_i dx_j}$. This gives us a matrix that has 2 along the diagonals, and zero everywhere else. Let $d = \dim(H)$. Let $x \in \mathbb{R}^d$.

$$\begin{aligned} x^T H x &= \\ x^T 2 I x &= \\ 2x^T I x &= \\ 2x^T x &= \\ 2||x||^2 &\geq 0 \end{aligned}$$

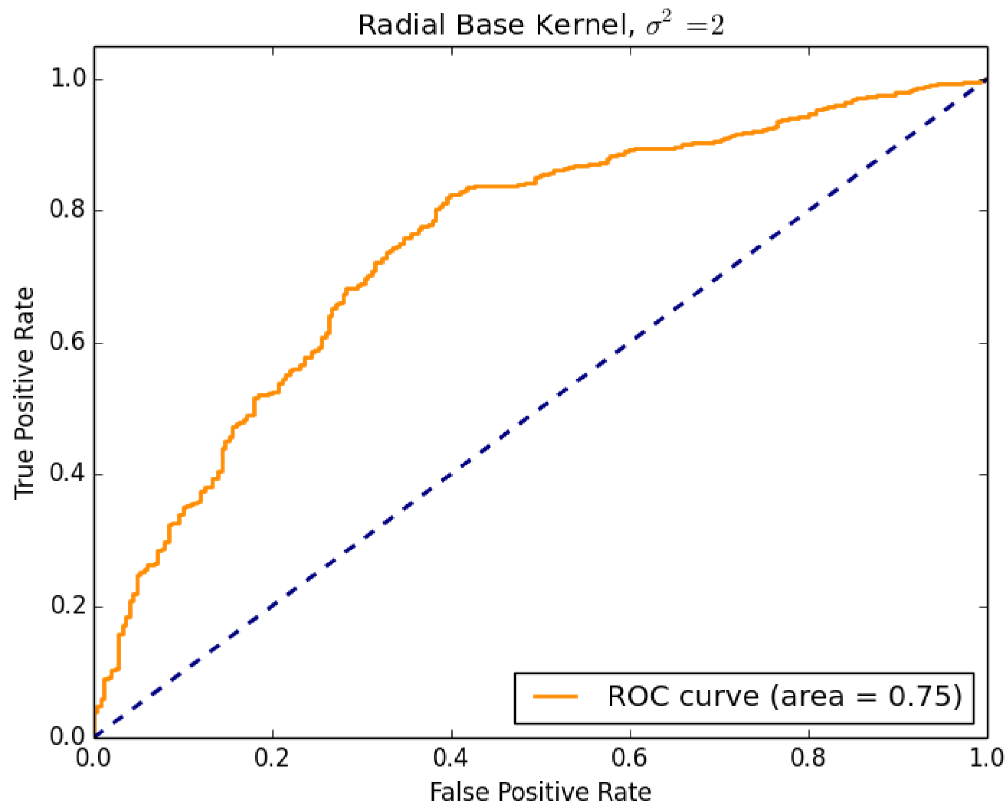
Since $x^T H x \geq 0 \forall x \in \mathbb{R}^d$, we H is positive semi-definite, and f is convex. Now, we just need to show that each g_i is convex – however, each g_i is a line. Since all lines are convex, we have that the g_i are convex, and therefore, the optimization problem is convex.

3. • a.



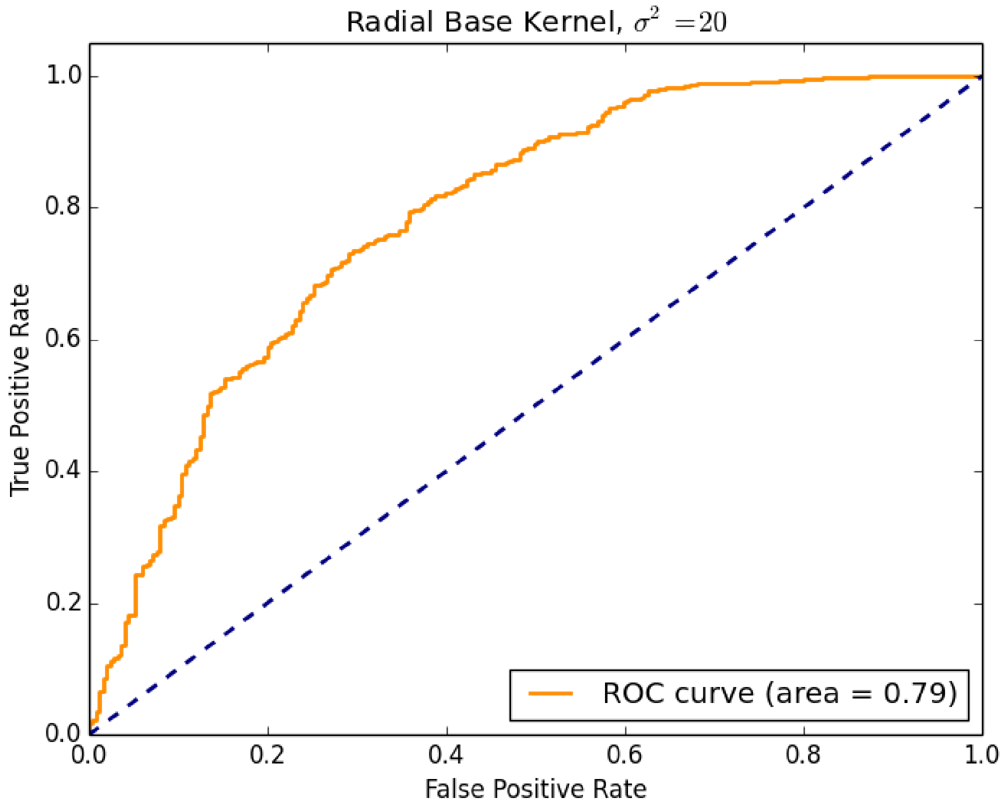
- b.

  Accuracy: 0.839



- c.

  Accuracy for $\sigma^2 = 2$: 0.785



Accuracy for $\sigma^2 = 20$: 0.836