1 Analytical (25 points)

1) Kernels (10 points) We say K is a kernel function if there exists some transformation $\phi : \mathbb{R}^m \to \mathbb{R}^{m'}$ such that $K(x, x') = \langle \phi(x), \phi(x') \rangle$. Let K_1 and K_2 be two kernel functions.

(a) Prove that $K(x, x') = K_1(x, x')K_2(x, x')$ is a kernel function.

Since K_1 and K_2 are kernel functions, $K_1(x, x') = f(x) * f(x')$ and $K_2(x, x') = g(x) * g(x')$ so then:
$K(x, x') = K_1(x, x')K_2(x, x')$
$= f(x) * f(x') * g(x) * g(x') = f(x) * g(x) * f(x') * g(x') = h(x) * h(x')$
where $h(x) = f(x) * g(x)$ and $h(x') = f(x') * g(x')$ and so that means $K(x, x')$ is also a kernel function since $K(x, x')$ is equal to $\phi(x) * \phi(x')$ for some $\phi = h$.

(b) Prove that $K(x, x') = K_1(x, x') + K_2(x, x')$ is a kernel function.

Since K_1 and K_2 are kernel functions, $K_1(x, x') = f(x) * f(x')$ and $K_2(x, x') = g(x) * g(x')$ so then:
$K(x, x') = K_1(x, x') + K_2(x, x')$
$= f(x) * f(x') + g(x) * g(x') = \phi(x) * \phi(x')$
where $\phi(x) = r(f(x), g(x))$ and $\phi(x') = r(f(x'), g(x'))$

2) Logistic Loss (10 points) Linear SVMs can be formulated in an unconstrained optimization problem

$$\min_{w,b} \sum_{i=1}^{n} H(y_i(w^T x_i)) + \lambda \|w\|_2^2, \tag{1}$$

where λ is the regularization parameter and H is the well known logistic loss function:

$$H(a) = \log(1 + exp(-a))$$

The logistic loss function can be viewed as a convex surrogate of the 0/1 loss function, which can be written using the indicator function as $I(a \leq 0)$.

(a) Prove that $H(a)$ is a convex function of a. $H(a) = \log(1 + exp(-a))$
$\frac{dH(a)}{da} = \frac{1}{1+exp(-a)} * exp(-a) * (-1)$
$= \frac{-exp(-a)}{1+exp(-a)}$
$\frac{d^2H(a)}{da^2} = \frac{(1+exp(-a))*exp(-a)++exp(-a)*(-exp(-a))}{(1+exp(-a))^2}$
$= \frac{exp(-a)}{(1+exp(-a))^2}$
and since $exp(-a)$ is always positive, then the numerator and denominator will always be positive.
Thus, the second derivative is always positive and so $H(a)$ is a convex function of a.

(b) The function $H(a) = \exp(-a)$ can also approximate the 0/1 loss function. How does this compare with the logistic loss function?

This exponential function has the same behavior for lim a $\to +\infty$, $-\infty$, and 0. Specifically, for a = 0, log(1 + exp(-a)) = log(1 + 1) = log(2) = 1 and exp(-a) = 1 so both functions have the same value of 1 at a = 0. Also for both, lim a $\to +\infty = 0$ and lim a $\to -\infty = +\infty$ One difference is that the exponential function is much steeper than the logistic function and has a larger slope value. However, both functions still divide the domain of a into two halfs: for $a < 0$, the function value is > 1, and for $a > 0$, the function value is < 1.

3) Margin (5 points) The SVM objective uses a margin value of 1 in the constraints ($\gamma = 1$.) Show that we can replace 1 with any arbitrary constant $\gamma > 0$ and that the solution for the maximum margin hyperplane is unchanged.

To solve for the maximum margin hyperplane, we write the quadratic program

$$argmin_w \frac{1}{2}||w||^2$$

s.t.

$$y_i(w \cdot x_i + b) - \gamma = 0, \forall i$$

- in the notes on svms, we did this with $\gamma = 1$

In order to solve this, we shall switch to the Dual Formulation and rewrite the objective using Lagrange multipliers. We add multipliers for each constraint: a_i s.t. $a_i \geq 0$. We have n of them. Thus we write the Lagrange function:

$$L(w, b, a) = \frac{1}{2}||w||^2 - \sum_{i=1}^{N} a_i y_i (w \cdot x_i + b) - \gamma$$

and this equation will be minimized when the norm of w is 0 and all the margins are one. So we take the derivative of L with respect to b:

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{n} a_i y_i$$

Setting that derivative to 0 gives the optimal solution for b which constrains the values of a. And we can already see that γ has not affected this constraint.

$$0 = -\sum_{i=1}^{n} a_i y_i$$

Then we take the derivative of L with respect to w for one of the positions of w:

$$\frac{\partial L}{\partial w_j} = w_j - \sum_{i=1}^{n} a_i y_i x_{i,j}$$

so we see that the value in w depends on values of x weighted by a which means that the full derivative is

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{n} a_i y_i x$$

and so we set the derivative equal to 0 and solve for w:

$$0 = w - \sum_{i=1}^{n} a_i y_i x$$

$$w = \sum_{i=1}^{n} a_i y_i x$$

and we can see that the value of γ did not affect this constraint. So this constraint is the same as for $\gamma = 1$ and so when we maximize this dual objective which minimizes our primal objective, we end up with the same solution regardless of the value of γ. This makes sense because the maximum separating hyperplane should still start in the same location. The only difference is that the margin is made bigger.