CMPUT 466/551 — Assignment 2

Instructors: R Greiner, B. Póczos
Due Date: 12:30pm, Tues, 3/Nov/09
The following exercises are intended to further your understanding of Linear Algebra (eigenvalues, ...), Dual Formulation, Lagrange Multiplier, Kernel Methods, Perceptrons, SVMs
Relevant reading: FTH: Chapters 5.8 and 12 (esp 12.3) + readings shown below.
Undergrads: solve problems 1–13
Grads: solve (all) problems 1–15
The `HW2-ReadMe.html` file describes the details of exactly what to hand in.
Total points: UGrad: 116 Grad: 162
[Hint: Several problems below extends results of previous problems...]

Question 1 *[12 points]* *Positive semi definite matrices*
The finite-dimensional spectral theorem says that any symmetric matrix $A \in \mathbb{R}^{n\times n}$ can be diagonalized by an orthogonal matrix. More explicitly: For every symmetric real matrix A there exists a real orthogonal matrix U such that $D = U^T AU \in \mathbb{R}^{n\times n}$ is a diagonal matrix. (Orthogonal means $U^T U = I$ where I is the identity matrix.) This matrix is "positive semi definite" (psd) iff $v^T Av \geq 0 \quad \forall v \in \mathbb{R}^n$.

a [4]: Use this theorem to prove that the eigenvalues of a symmetric matrix are real, and

b [4]: the eigenvectors $\{u^i\}$ are orthogonal (*i.e.*, $\langle u^i, u^j\rangle = 0$ when $i \neq j$).

c [4]: Let $A \in \mathbb{R}^{n\times n}$ be a symmetric matrix with eigenvalues $\{\lambda_1, \ldots, \lambda_n\}$. Prove that A is positive semi definite iff $\lambda_i \geq 0 \quad \forall i = 1, \ldots n$.

[Hint: $x^T x \geq 0$ for all $x \in \Re^n$. See also
http://en.wikipedia.org/wiki/Eigenvalue,_eigenvector_and_eigenspace
http://en.wikipedia.org/wiki/Spectral_theorem]

Question 2 *[4 points]* *Sum of positive semi definite matrices*
Assume that K_1, K_2 are positive semi definite matrices.

a [2]: Prove that, for any positive real constants $c_1,\ c_2 > 0$, $c_1K_1 + c_2K_2$ is psd.

b [2]: Prove that $K_1 - K_2$ is not necessarily psd.

Question 3 *[4 points]* *Constructing kernels*
Let $k_1(x, \tilde{x})$ and $k_2(x, \tilde{x})$ be valid kernel functions, and $c_1, c_2 > 0$ be positive real constants.

a [2]: Show that $c_1k_1(x, \tilde{x}) + c_2k_2(x, \tilde{x})$ is a valid kernel function, too.

b [2]: Show that $k_1 - k_2$ is not necessarily positive semi definite.

Question 4 *[12 points]* *Elementwise product of two positive semi definite matrices*
Let $K_1, K_2 \in \mathbb{R}^{n \times n}$ be two positive semi definite matrices. Prove that their **elementwise** product matrix $K(i,j) = K_1(i,j)K_2(i,j)$ is positive semi definite matrix, too.

[Hint: Consider combining two independent n-dimensional vectors $u = (u_1, \ldots, u_n)^T \sim N(0, K_1)$ and $v = (v_1, \ldots, v_n)^T \sim N(0, K_2)$, each drawn from its own Gaussian distribution.]

Question 5 *[4 points]* *Constructing kernels*
Let $k_1(x, \tilde{x})$ and $k_2(x, \tilde{x})$ be valid kernel functions. Show that $k_1(x, \tilde{x})k_2(x, \tilde{x})$ is also a valid kernel function.

Question 6 *[4 points]* *Product of positive semi definite matrices*
Let $A, B \in \mathbb{R}^{n \times n}$ be psd matrices.

a [2]: Show that AB is not necessarily positive semi definite.
[Hint: Does AB have to be symmetric?]

b [2]: Show that A^m is positive semi definite for all $m \in \mathbb{Z}_+$.

Question 7 *[2 points]* *Non kernel*
We know that $\exp(-\|x - y\|^2)$ is a kernel function. Show that

$$\exp(\|x - y\|^2)$$

is not a valid kernel.

Question 8 *[16 points]* *Perceptron* [Implementation]
a [6]: Describe when you expect the Primal to be faster than the Dual. ...and vice versa.

b [10]: Implement the perceptron classification algorithm in Primal and Dual form. Try to classify a $2D$ dataset, using "linear", "polynomial" and "RBF" kernels. The `HW2-ReadMe.html` file provides several datasets to play with — both linearly separable and non-separable cases. (It also specifies exactly what you should submit.)

Question 9 *[30 points]* *SVM* [Implementation]
Recall the primal problem for SVM is:

$$\begin{aligned} & \min_w \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \xi_i \\ & \text{subject to} \\ y_i\langle x_i, w\rangle \ \geq \ & 1 - \xi_i, \quad (i = 1, \ldots, m) \\ & \xi_i \geq 0, \quad (i = 1, \ldots, m) \end{aligned}$$

[[Correction (14/Oct): changed from ξ to ξ_i above.]]

a [6]: Show that this is the same as

$$\min_w \sum_{i=1}^{m} [1 - y_i\langle x_i, w\rangle]_+ \ + \ \lambda\|w\|^2$$

where in general $r_+ = \begin{cases} r & \text{if } r \geq 0 \\ 0 & \text{otherwise} \end{cases}$ is the positive part of r.

b [4]: Describe when you expect the Primal to be faster than the Dual. ...and vice versa.

c [20]: Implement the soft SVM classification problem in Primal and Dual form. **(You MAY use the 'quadprog' Matlab command... but may NOT use SVM toolboxes.)**

The `HW2-ReadMe.html` file provides a number of datasets. Compare the classification accuracy of your method using 'linear', 'polynomial(k)', and 'RBF' kernels. Feel free to play with the k and "C" parameters. The `HW2-ReadMe.html` file also specifies exactly what you should submit here.

Question 10 *[14 points] Constructing feature map in finite case* [Implement]
Let $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5\}$ consist of the following five 2D points:

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 2 \end{bmatrix}, \quad \begin{bmatrix} 2 \\ 3 \end{bmatrix}, \quad \begin{bmatrix} 3 \\ 1 \end{bmatrix}, \quad \begin{bmatrix} 1.5 \\ 1.5 \end{bmatrix}$$

a [2]: Plot these points using Matlab.

b [2]: Consider the kernel

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{8}) \quad 1 \leq i, j \leq 5$$

Use Matlab to show that its Gram matrix is positive semi definite, and thus that this $k(\cdot, \cdot)$ is a valid kernel.

c [6]: Using Matlab construct a feature map $\phi : \mathcal{X} \to \mathbb{R}^5$ that is compatible with kernel k.

d [4]: Verify if the constructed feature map is good — *i.e.*, if the inner product between $\phi(\mathbf{x}_i)$ and $\phi(\mathbf{x}_j)$ in the feature space is equal to the values of the kernel function $k(\mathbf{x}_i, \mathbf{x}_j)$.

Question 11 *[4 points]* l_p^{20} *norms* [Matlab exploration]
The `HW2-ReadMe.html` file provides three different 20 dimensional vectors x, each with only a few non-0 coordinates.

a [2]: For $p \in \{\frac{1}{128}, \frac{1}{32}, \frac{1}{2}, 1, 2, 8, 32, 128\}$, plot $\|x\|_p^p = \sum_{i=1}^{20} |x_i|^p$, and $\|x\|_p = (\sum_{i=1}^{20} |x_i|^p)^{1/p}$. You should probably use log-scale for the Y axis. (In Matlab: `set(gca,'Yscale','log')`.)

The `HW2-ReadMe.html` file specifies exactly what you should submit here.

b [2]: What happens when $p \to 0$? ...and when $p \to \infty$?

Question 12 *[4 points] Representer theorem*

a [2]: Let $\mathcal{F}$ be an RKHS function space with kernel $k(\cdot, \cdot)$. Let $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$ be m training input-output pairs. Our task is to find the $f^* \in \mathcal{F}$ function that minimizes the following regularized functional:

$$f^* = \arg\min_{f \in \mathcal{F}} \left(\prod_{i=1}^{m} |f(\mathbf{x}_i)|^6 \right) \sum_{i=1}^{m} \left[\left| \sin\left(\|\mathbf{x}_i\|^{|y_i - f(\mathbf{x}_i)|} \right) \right|^{25} + y_i |f(x_i)|^{42} \right] + \exp(\|f\|_{\mathcal{F}})$$

This is a nonparametric minimization problem over functions in the function space $\mathcal{F}$. Prove that f^* can be expressed as $f^*(\cdot) = \sum_{i=1}^{m} \alpha_i k(x_i, \cdot)$, reducing the problem to an m-dimensional minimization [with respect to $(\alpha_1, \ldots, \alpha_m)$] only.

b [2]: Now consider

$$g^* = \arg\min_{g \in \mathcal{F}} \; \|g\|_{\mathcal{F}} \sum_{i=1}^{m} \left[\left| \sin\left(\|\mathbf{x}_i\|^{|y_i - g(\mathbf{x}_i)|} \right) \right|^{25} + y_i |g(x_i)|^{42} \right] + \exp(\|g\|_{\mathcal{F}})$$

Can you use the representer theorem to express $g^*(\cdot) = \sum_{j=1}^{m} \alpha_j k(x_j, \cdot)$ for some α_j's? Explain.

Question 13 *[6 points] Lagrange multipliers, discrete random variables*
A discrete distribution $p = (p_1, p_2, \ldots, p_n)$ has $\sum p_i = 1$ and $p_i \geq 0$ for all i. The entropy, which measures the uncertainty of a distribution, is defined by $H(p) = -\sum_{i=1}^{n} p_i \log p_i$. (Note we define $0 \log 0 = 0$).

a [1]: Prove that the entropy is 0 for the ($p_1 = 1$, $p_2 = \ldots = p_n = 0$) deterministic distribution.

b [5]: Show that the uniform distribution has the largest entropy.

Question 14 *[16 points] Lagrange multipliers, continuous random variables* [Grad only]
The entropy of a continuous distribution with density function f is defined by $H(f) = -\int f(x) \log f(x)\, dx$. Let X be a random variable with density f.

a [8]: Prove that if $\mathbb{E}_f[X] = 0$ and $\mathbb{E}_f[X^2] = \sigma^2$, then the Gaussian distribution $N(0, \sigma^2)$ has the maximal entropy.

[Hint: Use Lagrange multipliers, and $\frac{\partial}{\partial f(y)} \int r(x) f(x)\, dx = r(y)$ *when* $r(.)$ *is not related to* $f(.)$. *(Note* $\frac{\partial}{\partial f(y)} \int f(x) \log(f(x))\, dx = \log(f(y)) + 1$.*)* `http://en.wikipedia.org/wiki/Functional_derivative`
Also, if a density has the form "$a \exp((x - b)^2 / 2c^2)$*" for any real constants* a, b *and* c*, then it must be the density of the normal distribution.]*

b [8]: Prove that if $support(f) = [0, \infty]$, and $E[X] = \mu$, then the exponential distribution $(f(x) = \frac{1}{\mu} \exp(-\frac{x}{\mu}))$ has the largest entropy.

[Hint: For support, see `http://en.wikipedia.org/wiki/Support_(mathematics)` *]*

Question 15 *[30 points] SVM, Quadratic Approximation, Dual form, Lagrange multipliers* [Grad only]
Given the following primal "**quadratic version**" of the soft SVM classification problem:

$$\min \frac{1}{2}\|w\|^2 \; + \; C \sum_{i=1}^{m} \xi_i^2$$

subject to

$$y_i \langle x_i, w \rangle \geq 1 - \xi, (i = 1, \ldots, m)$$
$$\xi \geq 0, (i = 1, \ldots, m)$$

What are the dual equations?