CSE 250B: Machine learning **Winter 2019**

Homework 4 — Convexity, Perceptron, SVM

1. Are the following functions $f : \mathbb{R} \to \mathbb{R}$ convex, concave, or neither? Justify your answer.
   (a) $f(x) = e^{ax}$, for some constant a.
   (b) $f(x) = |x|$.
   (c) $f(x) = \ln x$, where $x > 0$.
   (d) $f(x) = x^a$, for $a \geq 1$. What if $a \leq 0$? What if $0 \leq a \leq 1$?

2. Show that the following functions $f : \mathbb{R}^d \to \mathbb{R}$ are convex.
   (a) $f(x) = x^T M x$, where $M \in \mathbb{R}^{d \times d}$ is symmetric positive semidefinite.
   (b) $f(x) = e^{u \cdot x}$, for some $u \in \mathbb{R}^d$.
   (c) $f(x) = \max(f_1(x), \ldots, f_k(x))$, where $f_1, \ldots, f_k$ are convex.

3. Recall that the *entropy* of a discrete distribution $p = (p_1, \ldots, p_k)$ over k outcomes is defined as follows:

$$H(p) \;=\; \sum_{i=1}^{n} p_i \log \frac{1}{p_i}.$$

   Show that $H(p)$ is a concave function of p. You may switch to the natural logarithm if you wish.

4. Recall the loss function for regularized least squares: for some constant $\lambda > 0$,

$$L(w) = \sum_{i=1}^{n} (y^{(i)} - w \cdot x^{(i)})^2 + \lambda \|w\|^2,$$

   (a) Obtain an expression for the Hessian $H(w)$: that is, the $d \times d$ matrix of second derivatives.
   (b) Establish that $L(w)$ is a convex function of w.

5. In class, we studied convex *functions*. In this problem, we will define the notion of a convex *set*. Pick any $K \subseteq \mathbb{R}^d$. We say K is a convex set if for any $x, y \in K$, the line segment joining x and y lies entirely in K; more formally, for any $x, y \in K$ and any $0 < \theta < 1$, we have $\theta x + (1 - \theta) y \in K$.

   Which of the following is a convex set?
   (a) The circle: $\{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}$.
   (b) The unit ball: $\{x \in \mathbb{R}^d : \|x\| \leq 1\}$.
   (c) A hyperplane: $\{x \in \mathbb{R}^d : w \cdot x = 0\}$ for some $w \in \mathbb{R}^d$.
   (d) All k-sparse points: $\{x \in \mathbb{R}^d : x \text{ has at most } k \text{ nonzero coordinates}\}$.
   (e) The set of all $d \times d$ symmetric positive semidefinite matrices (treat each matrix as a vector in $\mathbb{R}^{d(d+1)/2}$).

6. *Norms.* In class, we talked about ℓ_p norms on $\mathbb{R}^d$, which include the following:

   - The l_1 norm: $\|x\|_1 = \sum_{i=1}^d |x_i|$.
   - The l_2 (Euclidean) norm: $\|x\| = \sqrt{\sum_{i=1}^d x_i^2}$.
   - The l_∞ norm: $\|x\|_\infty = \max_i |x_i|$.

   We now define norms more generally. A function $f : \mathbb{R}^d \to \mathbb{R}$ is a *norm* if:

   - It is nonnegative: $f(x) \geq 0$ always.
   - $f(x) = 0$ if and only if $x = 0$.
   - It is homogeneous: $f(tx) = |t|f(x)$ for any $x \in \mathbb{R}^d$ and $t \in \mathbb{R}$.
   - It satisfies the triangle inequality: $f(x + y) \leq f(x) + f(y)$.

   (a) Prove that the ℓ_1 norm satisfies these properties.

   (b) Prove that any norm $f : \mathbb{R}^d \to \mathbb{R}$ is a convex function. (This means we can easily incorporate norms into objective functions we are optimizing.)

   (c) Prove the following two properties. For the second, you may need to use the Cauchy-Schwarz inequality (that is, $|a \cdot b| \leq \|a\|\|b\|$ for any vectors a, b).

   - $\|x\|_1 \;\geq\; \|x\| \;\geq\; \|x\|_\infty$.
   - $\|x\|_1 \;\leq\; \|x\| \cdot \sqrt{d} \;\leq\; \|x\|_\infty \cdot d$.

   (d) Another norm that is quite common in machine learning and statistics is the Mahalanobis norm:

   $$\|x\|_A = \sqrt{x^T A x},$$

   where A is a symmetric positive definite matrix. What does the unit ball of this norm, that is $\{x : \|x\|_A \leq 1\}$, look like? *Hint*: think back to the multivariate Gaussian.

7. *A lower bound for the perceptron.* Give an example of a data set $\{(x^{(i)}, y^{(i)})\}$ for which the bound of the perceptron convergence theorem is tight. For convenience, choose the $x^{(i)}$ to have unit length, and show that the number of updates is $1/\gamma^2$.

8. *Small SVM example.* Consider the following small data set in $\mathbb{R}^2$:

   - Points $(1, 2), (2, 1), (2, 3), (3, 2)$ have label -1.
   - Points $(4, 5), (5, 4), (5, 6), (6, 5)$ have label $+1$.

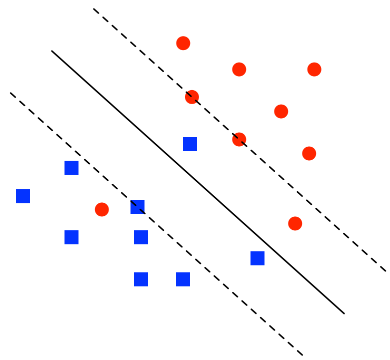
   Now, suppose (hard) SVM is run on this data.

   (a) Sketch the resulting decision boundary.

   (b) What is the (numerical value of the) margin, exactly?

   (c) What are w and b, exactly?

9. *Support vectors.* The picture below shows the decision boundary obtained upon running soft-margin SVM on a small data set of blue squares and red circles.



(a) Mark the support vectors. For each, indicate the approximate value of the corresponding slack variable.

(b) Suppose the factor C in the soft-margin SVM optimization problem were increased. Would you expect the margin to increase or decrease?