Homework IV

Due date: Monday Apr 3rd, 11:59pm

Problem 1: Convex Functions (40 points)

Prove the following functions are convex.

- (10') the log-sum-exponential function: $f(\mathbf{x}) = \log \sum_{i=1}^{n} \exp(x_i)$ is a convex function in terms of $\mathbf{x} = (x_1, \ldots, x_n)^\top$. (Hint: use the second order condition of convexity).

- (15') the objective of logistic regression for binary classification:

$$f(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(-y_i \mathbf{w}^\top \mathbf{x}_i)) + \frac{\lambda}{2} \|\mathbf{w}\|_2^2,$$

  where $\mathbf{x}_i \in \mathbb{R}^d$, $\mathbf{y}_i \in \{1, -1\}$. (Hint: first prove the convexity of the function $h(s) = \log(1 + \exp(s))$ and then use the rules that preserve convexity.)

- (15') the objective of support vector machine: $f(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} \max(0, 1 - y_i \mathbf{w}^\top \mathbf{x}_i) + \frac{\lambda}{2} \|\mathbf{w}\|_2^2$, where $\mathbf{x}_i \in \mathbb{R}^d$, $\mathbf{y}_i \in \{1, -1\}$. (Hint: similar as above).

Problem 2: the constrained version of Ridge Regression (30 points)

In class, we have mentioned the constrained version of ridge regression:

$$\begin{aligned} \min_{\mathbf{w} \in \mathbb{R}^d} \quad & \|\Phi \mathbf{w} - \mathbf{y}\|_2^2 \\ s.t. \quad & \|\mathbf{w}\|_2 \leq s \end{aligned}$$

where $\Phi \in \mathbb{R}^{n \times d}$ and $\mathbf{y} \in \mathbb{R}^n$. Answer the following questions:

- (20') Does strong duality hold? If yes, derive the KKT conditions regarding the optimal solution $\mathbf{w}_*$ to the above problem.

- (optional 10') Does a close-formed solution exist? If yes, derive the closed-form solution. If not, can you propose an algorithm for computing the optimal solution (describe the key steps of your algorithm)?

Problem 3: The equivalence between the Maximum Entropy Model and the Logistic Regression (40 points)

In class, we have talked about the maximum entropy model. For learning the posterior probabilities $\Pr(y|\mathbf{x}) = p(y|\mathbf{x})$ for $y = 1, \dots, K$ given a set of training examples $(\mathbf{x}_i, y_i), i = 1, \dots, n$, we can maximize the entropy of the posterior probabilities subject to a set of constraints, i.e.,

$$\begin{aligned}
\max_{p(y|\mathbf{x}_i)} \quad & -\sum_{i=1}^{n}\sum_{y=1}^{K} p(y|\mathbf{x}_i)\ln p(y|\mathbf{x}_i) \\
s.t. \quad & \sum_{y=1}^{K} p(y|\mathbf{x}_i) = 1 \\
& \sum_{i=1}^{n} \frac{\delta(y, y_i)}{n} f_j(\mathbf{x}_i) = \sum_{i=1}^{n} \frac{p(y|\mathbf{x}_i)}{n} f_j(\mathbf{x}_i), \quad j = 1, \dots, d, y = 1, \dots, K
\end{aligned}$$

where $\delta(y, y_i)$ is equal to 1 if $y_i = y$, and 0 otherwise, and $f_j(\mathbf{x}_i)$ is a feature function. Let us consider $f_j(\mathbf{x}_i) = [\mathbf{x}_i]_j$, i.e., the j-th coordinate of $\mathbf{x}_i$. Please show that the above Maximum Entropy Model is equivalent to the multi-class logistic regression model (without regularization). (Hint: use the Lagrangian dual theory)