CSCI 5525: Machine Learning (Fall 2020)

Exam 1

Due 10/22/2020 11:15 AM CDT

Exam Policy. This is a take-home exam so you are allowed to use the class books, lecture notes, and lecture videos. You are not allowed to discuss the exam with anyone else. Regarding online resources, you are not allowed to:

- Google around for solutions to exam problems,
- Ask for help on online,
- Look up things/post on sites like Quora, StackExchange, etc.

---

1. **(20 points)** In this problem, we consider generative and discriminative models.
   (a) **(10 points)** Explain the main assumptions in generative and discriminative classifiers, including how they make predictions on a test point.
   (b) **(10 points)** Consider a binary classification problem on a large dataset where you have limited knowledge of the process generating the data. Among logistic regression and naive Bayes, which classifier would you use and why?

2. (**15 points**) Given vectors $\mathbf{x}, \mathbf{z} \in \mathbb{R}^p$, answer True or False whether the following functions are valid kernels (True means it is a valid kernel, False means it is not a valid kernel):
   - $K(\mathbf{x}, \mathbf{z}) = \mathbf{x} - \mathbf{z}$ [True / False]
   - $K(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^\top \mathbf{z})^2$ [True / False]
   - $K(\mathbf{x}, \mathbf{z}) = -2\mathbf{x}^\top \mathbf{z}$ [True / False]

3. **(20 points)** This question considers first order methods, such as (sub)gradient descent and stochastic (sub)gradient descent, for solving binary classification problems using regularized logistic regression (RLR) of the form:

$$\min_{\mathbf{w} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \Big\{ -y_i \mathbf{w}^\top \mathbf{x}_i + \log(1 + \exp(\mathbf{w}^\top \mathbf{x}_i)) \Big\} + \lambda R(\mathbf{w}) \ ,$$

   where n is the number of data points, $R : \mathbb{R}^p \to \mathbb{R}$ is a convex, possibly non-smooth, regularization function, $\lambda \in \mathbb{R}_+$ is the regularization parameter, $\mathbf{x}_i \in \mathbb{R}^p$ is a feature vector, and $y_i \in \{0, 1\}$ is a class label for all i.

(a) **(10 points)** In general, why would you use stochastic gradient descent over gradient descent? List at least two reasons.

(b) **(5 points)** Let $R(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 = \frac{1}{2}\sum_{j=1}^{p} w_j^2$. In order to get ϵ-close to the global objective[1], what are the total runtimes of gradient descent and stochastic gradient descent for the RLR problem in terms of (n, ϵ).

(c) **(5 points)** Let $R(\mathbf{w}) = \|\mathbf{w}\|_1 = \sum_{j=1}^{p} |w_j|$. In order to get ϵ-close to the global objective, what are the total runtimes of sub-gradient descent and stochastic sub-gradient descent algorithms for the RLR problem in terms of (n, ϵ).

4. **(45 points)** In this problem, we consider the perceptron algorithm which can be viewed as applying stochastic gradient descent to a non-smooth convex objective function.

(a) **(10 points)** Write down the objective function and explain why it is non-smooth.

(b) **(25 points)** Consider starting the perceptron algorithm from initial vector $\mathbf{w}_0 = 0$, and let the final vector after convergence be $\mathbf{w}^* = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$. Show that α_i is exactly the number of mistakes made on $(\mathbf{x}_i, y_i)$ during the perceptron iterations before convergence.

(c) **(10 points)** If the classes are not linearly separable, argue why the perceptron with a fixed learning rate η may not converge. Can you modify the learning rate η so that the iterations converge in expectation? Explain your answer.

[1]For stochastic gradient descent, we consider the expected performance.

Extra Credit Problems. *In order for extra credit solutions to be considered, you must provide reasonable solutions to all parts of Problems 1-4 above. If you skip any part of a problem or do not provide a reasonable solution (i.e., make a real effort), we will not count any extra credit towards your grade.*

1. **Extra Credit 1. (5 points)** In this problem, we consider maximum likelihood estimation. Let $x_1, x_2, ..., x_n$ be i.i.d. samples drawn from the following distribution:

   $$P_\theta(x) = 2\theta x e^{-\theta x^2}$$

   where θ is the parameter and x is a positive real number. Derive the maximum likelihood estimate for θ. (For full credit you must show all steps and not leave the final solution as an optimization problem.)

2. **Extra Credit 2. (5 points)** For any $a \in \mathbb{R}$, let $\sigma(a) = \frac{1}{1+\exp(-a)}$. For each sample $(\mathbf{x}_i, y_i) \in \mathbb{R}^p \times \{0, 1\}$, the conditional log-likelihood of logistic regression is

   $$\ell(y_i \mid \mathbf{x}_i, \mathbf{w}) = y_i \log(\sigma(\mathbf{w}^\top \mathbf{x}_i)) + (1 - y_i) \log(\sigma(-\mathbf{w}^\top \mathbf{x}_i)).$$

   Derive the gradient of $\ell(y_i \mid \mathbf{x}_i, \mathbf{w})$ with respect to w_j (the j-th coordinate of $\mathbf{w}$), i.e., $\frac{\partial}{\partial w_j} \ell(y_i \mid \mathbf{x}_i, \mathbf{w})$. (You must clearly mention the derivative properties used.)

3. **Extra Credit 3. (5 points)** Let N be any positive integer. For every $x, y \in \{1, \ldots, N\}$ define $K(x, y) = \min\{x, y\}$. Show that K is a valid kernel.