CSCI 5525: Machine Learning (Fall'15)

Take Home Final: Due 12/18/15, 11:55 pm

1. **(20 points)** We consider logistic regression for a 2-class classification setting. Let $D = \{(\mathbf{x}_i, y_i)_{i=1}^n\}$ be a dataset for 2-class classification, where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{0, 1\}$.

   (a) (10 points) Assume the class conditional distributions are members of an exponential family distribution, so that

   $$p(\mathbf{x}|C_1) = \exp(\eta_1^T \mathbf{x}) g(\eta_1) h(\mathbf{x}) \ , \quad \text{and} \quad p(\mathbf{x}|C_2) = \exp(\eta_2^T \mathbf{x}) g(\eta_2) h(\mathbf{x}) \ . \tag{1}$$

   Show that the log-odds over the class posterior distributions is affine, which is the assumption logistic regression makes, i.e.,

   $$\log \left(\frac{p(C_1|\mathbf{x})}{p(C_2|\mathbf{x})} \right) = \mathbf{w}^T \mathbf{x} + w_0 \ , \tag{2}$$

   for suitable parameters $(\mathbf{w}, w_0)$.

   (b) (10 points) The sparse logistic regression formulation considers a L_1-norm regularized version of standard logistic regression, and the loss function is given by:

   $$E(\mathbf{w}) = \sum_{i=1}^{n} \{-y_i \langle \mathbf{w}, \mathbf{x}_i \rangle + \log(1 + \exp(\langle \mathbf{w}, \mathbf{x}_i \rangle))\} + \lambda \|\mathbf{w}\|_1 \ , \tag{3}$$

   where $\|\mathbf{w}\|_1 = \sum_{i=1}^{d} |w_j|$. Show that $E(\mathbf{w})$ is a convex function of $\mathbf{w}$.

2. **(30 points)** Consider the problem of solving a constrained quadratic optimization problem over $\mathbf{x} \in \mathbb{R}^d$ given as follows:

   $$\min_{\mathbf{x}} \ \mathbf{x}^T P \mathbf{x} + q^T \mathbf{x} \quad \text{such that} \quad A\mathbf{x} = \mathbf{a}, B\mathbf{x} \leq \mathbf{b} \ ,$$

   where $P \in \mathbb{R}^{d \times d}$ is positive definite, $q \in \mathbb{R}^d$, $A \in \mathbb{R}^{k_1 \times d}$, $\mathbf{a} \in \mathbb{R}^{k_1}$, $B \in \mathbb{R}^{k_2 \times d}$, and $\mathbf{b} \in \mathbb{R}^{k_2}$.

   (a) (8 points) What is the Lagrangian for the optimization problem? Clearly state the dimensionality and constraints on the Lagrange multipliers involved.

   (b) (10 points) Can the Lagrange dual for the optimization problem be written as a closed form expression in terms of the Lagrange multipliers and the given matrices and vectors? If yes, clearly state the closed form for the Lagrange dual; if no, clearly explain why a closed form is not possible.

   (c) (12 points) Clearly present an ADMM algorithm for solving the optimization problem. In particular, say how additional variables (if any) need to be introduced, and then give the key steps for the algorithm. The key steps have to be presented as closed form formulae, not as minimization over functions.

3. **(25 points)** Consider the problem of learning a mixture of Gaussians using the EM (expectation maximization) algorithm. We assume that there are K Gaussians, and all of them have the same covariance Σ, i.e., the component Gaussians are of the form $N(\boldsymbol{\mu}_k, \Sigma)$. Let π_k denote the prior probability of each component, so that $\pi_k \geq 0$ and $\sum_{k=1}^{K} \pi_k = 1$. Assume a set of N samples $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ drawn independently from the mixture of Gaussians.

   (a) (10 points) Give a closed form expression for the E-step, where we compute the posterior probability $p(z_k|\mathbf{x}_n)$. Simplify the expression by using the fact that all the Gaussians have the same covariance.

   (b) (10 points) Give closed form expressions for the M-step, where we compute $\pi_k, \boldsymbol{\mu}_k, k = 1, \ldots, K$ and Σ.

   (c) (5 points) Does the computation of Σ get simplified since it is the same for all Gaussians? Clearly explain your answer.

4. **(25 points)** Consider the setting of non-parametric regression using Gaussian processes (GPs). Let $(X, \mathbf{y})$ denotes the training set of n points (i.e., $X \in \mathbb{R}^{n \times d}$ and $\mathbf{y} \in \mathbb{R}^n$), K denote the kernel function, and X^* denote the test set of m points, i.e., $X^* \in \mathbb{R}^{m \times d}$.

   (a) (12 points) Given the expression for the mean and variance of the predictive distribution $p(y^*|x^*, X, \mathbf{y})$ for any single test point x^*. Using the expression, show that the mean can be expressed in a linear form in two different ways.

   (b) (7 points) Let a test point x^* be exactly the same as a training point x_i, with training value y_i. Does the mean of the predictive distribution $p(y^*|x^*, X, \mathbf{y})$ at x^* exactly overlap with y_i? Clearly explain your answer.

   (c) (6 points) Consider the setting with m test points X^*, so that the joint predictive distribution will be a m-dimensional distribution $p(\mathbf{y}^*|X^*, X, \mathbf{y})$. Do you think that one can obtain a closed form expression for the m-dimensional predictive joint distribution? Clearly explain your answer. Please note that the question is not asking you to try to derive the closed form if it exists—rather, present an argument based on your understanding of GPs.