Machine Learning Final Exam, Spring 2019

姓名:_____ 学号:_____ 总分:_____

2019 年 5 月 23 日

1 Logistic regression and MLE (10 + 5 = 15 points)

Suppose we have training data set $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$ with binary label $y_i \in \{0, 1\}$ and that given $\mathbf{x}$ the label y satisfies a Bernoulli distribution

$$P(y \mid \mathbf{x}) = \begin{cases} \frac{e^{\mathbf{w}_0^T \mathbf{x}}}{e^{\mathbf{w}_0^T \mathbf{x}} + e^{\mathbf{w}_1^T \mathbf{x}}}, & \text{if } y = 0, \\ \frac{e^{\mathbf{w}_1^T \mathbf{x}}}{e^{\mathbf{w}_0^T \mathbf{x}} + e^{\mathbf{w}_1^T \mathbf{x}}}, & \text{if } y = 1. \end{cases}$$

(a) Now use maximum likelihood estimation to derive the objective function of the Logistic regression, i.e., the loss function.

(b) Suppose you obtain the optimal $\mathbf{w}^* = [\mathbf{w}_0^*; \mathbf{w}_1^*]$ from training. How do you use $\mathbf{w}^*$ to make prediction?

2 Learning theory ($5 \times 5 = 25$ points)

(a) Suppose hypothesis set $\mathcal{H}$ has break point = 2, what is the possible largest value for $m_{\mathcal{H}}(3)$? Here $m_{\mathcal{H}}$ is the growth function.

(b) Give your understanding of what VC dimension is.

(c) What is the VC dimension of the (zero-centered) circle? Give your reasons.

$$\mathcal{H} = \{h_\theta(x) = \text{sign}(\|x\|_2^2 + \theta), \theta \in \mathbb{R}_+, x \in \mathbb{R}^2\}.$$

(d) What is generalization and overfitting? Give your understanding.

(e) Use Vapnik-Chervonenkis inequality

$$\mathbb{P}[|\mathbb{E}_{in}(g) - \mathbb{E}_{out}(g)| > \epsilon] \leq 4m_{\mathcal{H}}(2N)e^{-\frac{1}{8}\epsilon^2 N}$$

to analyze which factors can lead to bad generalization? What is the technique of detecting overfitting? How would you fix overfitting if it happens?

(Here $\mathcal{H}$ is the hypothesis set, $g \in \mathcal{H}$ is the hypothesis with optimal in-sample error, $\mathbb{E}_{in}$ denotes the in-sample error, N is the sample size, and $\mathbb{E}_{out}$ denotes the out-of-sample error.)

3 Regularization ($5 + 5 + 5 = 15$ points)

Suppose we use the ℓ_1 regularization problem as follows:

$$\min_{\theta} \|\theta\|_1 \quad \text{s.t. } \frac{1}{2}\|X\theta - y\|_2^2 \leq \epsilon.$$

for our future prediction. Here $X \in \mathbb{R}^{m \times n}$ is the historical data set matrix, $\theta \in \mathbb{R}^n$, observation $y \in \mathbb{R}^m$ and $\epsilon \geq 0$ is the regularization parameter.

(a) Analyze how ϵ affects the regularization effect.

(b) If you were using this formulation for your regularization, what is the value range of ϵ you use?

(c) Suppose $\theta \in \mathbb{R}^4$. Which one of the following graphs should be the typical path plot for this regularization? Give your reasons.

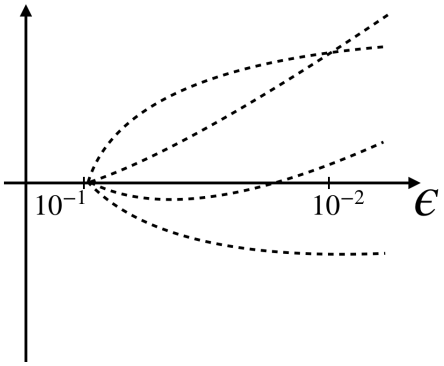

Figure 1

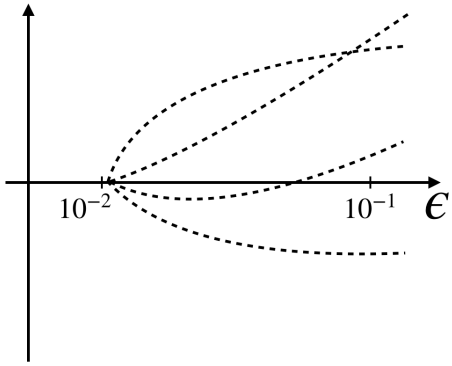

Figure 2

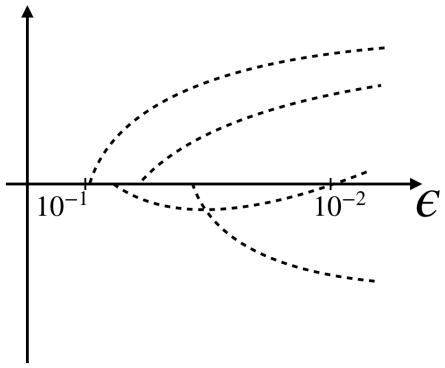

Figure 3

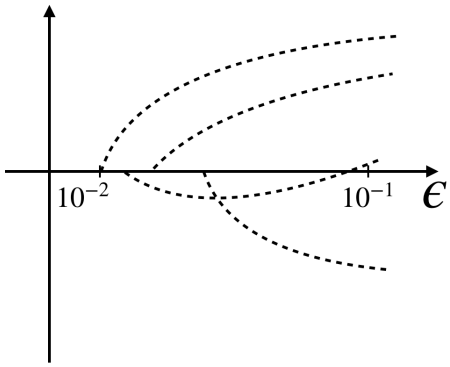

Figure 4

4 Second-order methods (5×5 = 25 points)

Consider the following algorithm for minimizing smooth objective $f(x)$:

$$x_{k+1} \leftarrow x_k - H_k \nabla f(x_k),$$

where $H_k \in \mathbb{R}^{n \times n}$ is positive definite.

(a) Derive the local model $m_k(x)$ to approximate f at x_k, so that x_{k+1} is the optimal solution of $m_k(x)$.

(b) Show that $d_k = x_{k+1} - x_k$ is a descent direction at x_k.

(c) Write down the secant equation for a quasi-Newton method to satisfy.

(d) Suppose f is convex, and we want to use $H_k = \alpha_k I$ in the local model. Notice that the curvature condition may not be satisfied, since we have a system of n equations with

one variable. Therefore, we can choose α_k as the least squares solution to the secant equation. Write down the formulation of α_k.

(e) Suppose f is convex, show that the α_k you found in (d) is nonnegative, i.e., $\alpha_k \geq 0$.

5 Stochastic gradient descent ($4 + 3 + 4 + 9 = 20$ points)

Consider a stochastic gradient method for solving $\min_w F(w)$:

$$w_{k+1} \leftarrow w_k - \alpha_k g(w_k, \xi_k)$$

where $g(w_k, \xi_k)$ is a stochastic gradient at w_k—i.e., an estimator of $\nabla F(w_k)$ and ξ_k represents a seed for generating a stochastic direction.

(a) Suppose F is Lipschitz differentiable with Lipschitz constant $L > 0$. Show that

$$F(w_{k+1}) \leq F(w_k) + \nabla F(w_k)^T (w_{k+1} - w_k) + \frac{L}{2} \|w_{k+1} - w_k\|_2^2.$$

(b) Using the result in (a), show that at x_k:

$$\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) \leq -\alpha_k \nabla F(w_k)^T \mathbb{E}_{\xi_k}[g(w_k, \xi_k)] + \frac{1}{2} \alpha_k^2 L \mathbb{E}_{\xi_k}[\|g(w_k, \xi_k)\|_2^2].$$

(c) Suppose $g(w_k, \xi_k)$ is an unbiased estimator of $\nabla F(w_k)$ with finite variance, use the result in (b) to show

$$\mathbb{E}_{\xi_k}[F(w_{k+1})] - F(w_k) \leq (\frac{L}{2} \alpha_k^2 - \alpha_k) \|\nabla F(w_k)\|_2^2 + \frac{L}{2} \alpha_k^2 \mathrm{Var}_{\xi_k}[g(w_k, \xi_k)].$$

(d) Use the inequality derived in (c), discuss how you would choose learning rate α_k to enforce convergence of SGD. (assume the variance of $g(w_k, \xi_k)$ is fixed for each iteration, you don't have to prove the explicit condition of guaranteeing convergence)

姓名: 学号:

答题纸

姓名: 学号:

答题纸

姓名: 学号:

答题纸

姓名: 学号:

答题纸

姓名: 学号:

答题纸

姓名: 学号:

答题纸

姓名: 学号:

答题纸