10702/36702 Statistical Machine Learning, Spring 2008: Midterm Solutions

March 17, 2008

1 Regression [25 points] (Robin)

Let $X_1 \in \mathbb{R}$ and $X_2 \in \mathbb{R}$ and

$$Y = m(X_1, X_2) + \epsilon \tag{1}$$

where $\mathbb{E}(\epsilon) = 0$.

(a) Consider the class of multiplicative predictors of the form $m(x_1, x_2) = \beta x_1 x_2$. Let β_* be the best predictor, that is, β_* minimizes $\mathbb{E}(Y - \beta X_1 X_2)^2$. Find an expression for β_*.

★ **SOLUTION:** $R = E(Y - \beta X_1 X_2)^2$
$\frac{\partial R}{\partial \beta} = -2E(Y - \beta X_1 X_2)X_1 X_2 = 0$
$\Rightarrow \beta_* = \frac{E(YX_1X_2)}{E(X_1^2 X_2^2)}$

(b) Suppose the true regression function is

$$Y = X_1 + X_2 + \epsilon.$$

Also assume that $\mathbb{E}(X_1) = \mathbb{E}(X_2) = 0$, $\mathbb{E}(X_1^2) = \mathbb{E}(X_2^2) = 1$ and that X_1 and X_2 are independent. Find the predictive risk $R = \mathbb{E}(Y - \beta_* X_1 X_2)^2$ where β_* was defined in part (a).

★ **SOLUTION:**

$$\begin{aligned}
\beta_* &= \frac{E(YX_1X_2)}{E(X_1^2)E(X_2^2)} = E(YX_1X_2) \\
&= E((X_1 + X_2 + \epsilon)(X_1X_2)) \\
&= E(X_1^2 X_2 + X_1 X_2^2 + X_1 X_2) \\
&= 0 \\
\text{Hence,} \quad E(Y - \beta X_1 X_2)^2 &= E(Y^2) \\
&= E((X_1 + X_2 + \epsilon)^2) \\
&= E(X_1^2 + X_2^2 + \epsilon^2 + 2X_1X_2 + 2X_1\epsilon + 2X_2\epsilon) \\
&= 2 + E(\epsilon^2)
\end{aligned}$$

(c) We are given n observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ from (1). Give an estimator $\hat{\beta}_n$ for β_* and show that it is consistent.

★ **SOLUTION:** $\hat{\beta} = \frac{\frac{1}{n}\sum Y_i X_{1i} X_{2i}}{\frac{1}{n}\sum X_{1i}^2 X_{2i}^2}$
$\frac{1}{n}\sum Y_i X_{1i} X_{2i} \xrightarrow{p} E(YX_1X_2)$ $\quad \frac{1}{n}\sum X_{1i}^2 X_{2i}^2 \xrightarrow{p} E(X_1^2 X_2^2)$ $\quad \therefore \hat{\beta} \to \beta$

2 Bayes and Minimax [25 points] (Jingrui)

Let $X_1, \ldots, X_n \sim f(x;\theta)$ where $f(x;\theta)$ is a distribution from the family of distributions

$$\mathcal{P} = \{f(x;\theta) : \theta \in \Theta\}.$$

Let the loss function for an estimator $\hat{\theta}$ be

$$L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$$

(a) Define the risk function $R(\theta, \hat{\theta})$.

★ **SOLUTION:**

$$R(\theta, \hat{\theta}) = E[L(\theta, \hat{\theta})]$$

(b) Define the minimax estimator.

★ **SOLUTION:** $\hat{\theta}$ minimizes $\sup_\theta R(\theta, \hat{\theta})$.

(c) Let $\pi(\theta)$ denote a prior distribution. Define the Bayes' estimator $\hat{\theta}_\pi$ with respect to π.

★ **SOLUTION:** $\hat{\theta}_\pi$ minimizes $R_\pi = \int R(\theta, \hat{\theta})\pi(\theta) d\theta$.

(d) Show that the Bayes estimator is

$$\hat{\theta}_\pi = \mathbb{E}(\theta | X_1, \ldots, X_n).$$

★ **SOLUTION:** $R_\pi = \int [\int (\theta - \hat{\theta})^2 f(\theta | X_1 = x_1, \ldots, X_n = x_n) d\theta] m(x_1, \ldots, x_n) dx_1 \ldots, dx_n$. Taking the partial derivative of $\int (\theta - \hat{\theta})^2 f(\theta | x_1, \ldots, x_n) d\theta$ with respect to $\hat{\theta}$, we have

$$\frac{\partial}{\partial \hat{\theta}} \int (\theta - \hat{\theta})^2 f(\theta | x_1, \ldots, x_n) d\theta = 2 \int (\hat{\theta} - \theta) f(\theta | x_1, \ldots, x + n) d\theta$$

Setting it to 0, we get $\hat{\theta} = \int \theta f(\theta | x_1, \ldots, x_n) d\theta = \mathbb{E}(\theta | X_1, \ldots, X_n)$.

(e) Suppose that $R(\theta, \hat{\theta}_\pi) = c$ for some constant c. Show that $\hat{\theta}_\pi$ is minimax.

★ **SOLUTION:** Let $\tilde{\theta}$ be any other estimator, then

$$\sup_\theta R(\theta, \tilde{\theta}) \geq \int R(\theta, \tilde{\theta}) \pi(\theta) d\theta \geq \int R(\theta, \hat{\theta}_\pi) \pi(\theta) d\theta = c = \sup_\theta R(\theta, \hat{\theta}_\pi)$$

Therefore, $\hat{\theta}_\pi$ is minimax.

3 Model Selection [25 points] (Robin)

Suppose we have the following data: $(X_1, Y_1), \ldots, (X_n, Y_n)$ where $Y_i \in \mathbb{R}$ and $X_i \in \mathbb{R}^p$. Assume that $p < n$. Also assume that

$$Y_i = X_i^T \beta + \epsilon_i$$

where ϵ_i has mean 0. Let $\mathbb{X}$ be the $n \times p$ design matrix, that is, $\mathbb{X}(i,j) = X_{ij}$. Suppose that $\mathbb{X}^T\mathbb{X} = I$ where I is the $p \times p$ identity matrix. (We say that the design matrix is orthogonal.)

(a) Recall that the ridge regression estimator is

$$\hat{\beta} = (\mathbb{X}^T\mathbb{X} + \lambda I)^{-1}\mathbb{X}^T Y$$

where $Y = (Y_1, \ldots, Y_n)^T$. Find the predictive risk of $\hat{m}(x) = x^T\hat{\beta}$. Hint: first find the mean and variance of $\hat{\beta}$.

★ **SOLUTION:** $\hat{\beta} = (X^TX + \lambda I)^{-1}X^TY = (I + \lambda I)^{-1}X^TY = \frac{1}{1+\lambda}X^TY = \frac{1}{1+\lambda}X^T[X\beta + \epsilon] = \frac{\beta}{1+\lambda} + \frac{1}{1+\lambda}X^T\epsilon$
$\bar{\beta} = E(\hat{\beta}) = \frac{\beta}{1+\lambda}$ $\quad V(\hat{\beta}|X) = \frac{\sigma^2}{(1+\lambda)^2}X^TX = \frac{\sigma^2}{(1+\lambda)^2}I$
Also, $\beta - \bar{\beta} = \frac{\lambda}{1+\lambda}\beta$
$S = V(\hat{\beta}|X) = (\frac{\sigma}{(1+\lambda)})^2 I \quad R = E(Y - X^T\hat{\beta})^2$

$$\begin{aligned}
E(Y - X^T\hat{\beta})^2 &= E(X\beta + \epsilon - X^T\hat{\beta})^2 \\
&= E[(\hat{\beta} - \beta)^T XX^T(\hat{\beta} - \beta)] + \sigma^2 \\
&= E[(\hat{\beta} - \bar{\beta})^T XX^T(\hat{\beta} - \bar{\beta})] + 2E[(\hat{\beta} - \bar{\beta})^T XX^T(\bar{\beta} - \beta)] + E[(\bar{\beta} - \beta)^T XX^T(\bar{\beta} - \beta)] + \sigma^2 \\
&= \sum_{j=1}^{p} E(X_j^2)[(\frac{\lambda}{1+\lambda})^2\beta_j^2 + (\frac{\sigma}{1+\lambda})^2] + 0 + (\frac{\lambda}{1+\lambda})^2\beta^T E(XX^T)\beta + \sigma^2
\end{aligned}$$

(b) Still assuming that the design matrix is orthogonal, show that it is possible to find the lasso estimator without using iterative algorithms or quadratic programming. Hint: consider the transformed response $Z = \mathbb{X}^TY$.

★ **SOLUTION:** $Z = X^TY = X^T(X\beta + \epsilon) = \beta + X^T\epsilon$
$Z \sim N(\beta, \sigma^2)$
Apply soft thresholding to Z

4 Convex Duality [25 points] (Jingrui)

Let $X_i \sim \text{Bernoulli}(\theta)$ be independent, with observations $\{X_1, X_2, X_3\} = \{0, 1, 0\}$. Thus, $\mathbb{P}(X_i = 1) = \theta$ and $\mathbb{P}(X_i = 0) = 1 - \theta$ where $0 \le \theta \le 1$. Consider the optimization problem

$$\min_{\theta} \; f(\theta)$$
$$\text{such that } \theta \ge 1/2$$

where $f(\theta)$ is the negative log-likelihood.

(a) What is the solution to this problem?

★ **SOLUTION:** The likelihood is $L = \theta(1-\theta)^2$. Therefore, $f(\theta) = -\log\theta - 2\log(1-\theta)$, which is a convex function. Let $\frac{\partial f(\theta)}{\partial\theta} = 0$, we get $\hat{\theta} = 1/3$. However, this solution does not satisfy the constraint. When $\theta \geq 1/2$, $f(\theta)$ is a decreasing function. Therefore, the solution to this problem is $\hat{\theta} = 1/2$.

(b) Write the Lagrangian.

★ **SOLUTION:**

$$L(\theta,\lambda) = -\log\theta - 2\log(1-\theta) + \lambda(\frac{1}{2} - \theta)$$

(c) Derive the dual problem.

★ **SOLUTION:** $\frac{\partial L(\theta,\lambda)}{\partial\theta} = -\frac{1}{\theta} + \frac{2}{1-\theta} - \lambda = 0$. Therefore, $\lambda\theta^2 + (3-\lambda)\theta - 1 = 0$, and $\theta^* = \frac{\lambda-3+\sqrt{(\lambda-3)^2+4\lambda}}{2\lambda}$. The dual function: $l(\lambda) = -\log\theta^* - 2\log(1-\theta) + \lambda(\frac{1}{2} - \theta^*)$.

(d) State the KKT conditions.

★ **SOLUTION:**

$$\begin{aligned}
-\frac{1}{\theta^*} + \frac{2}{1-\theta^*} - \lambda^* &= 0 \\
\frac{1}{2} - \theta^* &\leq 0 \\
\lambda^* &\geq 0 \\
\lambda^*(\frac{1}{2} - \theta^*) &= 0
\end{aligned}$$

5 Regularization [25 points] (Robin)

Let Y be the random variable

$$Y = \mu + \epsilon$$

where $\epsilon \sim N(0,1)$ and $\mu \in \mathbb{R}$ in a constant. The elastic net estimator $\hat{\mu}$ is defined to be the value of μ that minimizes

$$M(\mu) = (Y-\mu)^2 + \lambda|\mu| + \alpha\mu^2$$

where λ, $\alpha > 0$. Find $\hat{\mu}$.

★ **SOLUTION:** $\frac{\partial M}{\partial\mu} = -2(Y-\mu) + \lambda z + 2\alpha\mu$

where $z = \begin{cases} 1 & \text{if } \mu > 0 \\ -1 & \text{if } \mu < 0 \\ \in [-1,1] & \text{if } \mu = 0 \end{cases}$

When $\mu = 0$, $-2Y + \lambda z = 0 \quad Y = \frac{\lambda z}{2} \quad \therefore \hat{\mu} = 0 \quad \text{if } |Y| \leq \frac{\lambda}{2}$

When $\mu > 0$, $-2(Y-\mu) + \lambda + 2\alpha\mu = 0 \therefore \hat{\mu} = \frac{2Y-\lambda}{2(1+\alpha)}$

When $\mu < 0$, $-2(Y-\mu) - \lambda + 2\alpha\mu = 0 \therefore \hat{\mu} = \frac{2Y+\lambda}{2(1+\alpha)}$

$$\hat{\mu} = \begin{cases} \frac{2Y-\lambda}{2(1+\alpha)} & Y > \lambda/2 \\ 0 & |Y| \leq \lambda/2 \\ \frac{2Y+\lambda}{2(1+\alpha)} & Y < -\lambda/2 \end{cases}$$

6 Mixture Models [25 points] (Jingrui)

Let $(Z_1, Y_1), \ldots, (Z_n, Y_n)$ be generated as follows:

$$Z_i \sim \text{Bernoulli}(p)$$
$$Y_i \sim \begin{cases} N(0,1) & \text{if } Z_i = 0 \\ N(5,1) & \text{if } Z_i = 1 \end{cases}$$

(a) Assume we do not observe the Z_i's. Write the distribution $f(y)$ of Y as a mixture.

★ **SOLUTION:**

$$f(y) = p\phi(y-5) + (1-p)\phi(y)$$

where $\phi(\cdot)$ is the pdf of a standard normal distribution.

(b) Write down the likelihood function for p.

★ **SOLUTION:**

$$L(p) = \prod_{i=1}^{n} [p\phi(y_i - 5) + (1-p)\phi(y_i)]$$

(c) Write down the complete likelihood function for p (assuming the Z_i's are observed).

★ **SOLUTION:**

$$L(p) = \prod_{i=1}^{n} [p^{z_i}(\phi(y_i - 5))^{z_i}(1-p)^{1-z_i}(\phi(y_i))^{1-z_i}]$$

(d) Find a consistent estimator of p that avoids using EM.

★ **SOLUTION:** $\mathbb{E}(Y) = 5p + 0(1-p) = 5p$. Let $\hat{p} = \frac{\bar{Y}}{5}$. $\mathbb{E}(\hat{p}) = p$. According to Law of Large Numbers, $\hat{p}$ converges to $\mathbb{E}(\hat{p})$ in probability. Therefore, $\hat{p}$ is a consistent estimator of p.

7 Classification [25 points] (Robin)

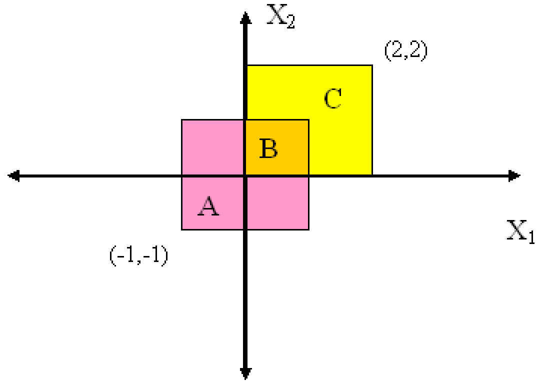
Suppose that $\mathbb{P}(Y = 1) = \mathbb{P}(Y = 0) = \frac{1}{2}$ and

$$X|Y = 0 \sim \text{Uniform on } S_0$$

$$X|Y = 1 \sim \text{Uniform on } S_1$$

where S_0 is the square in $\mathbb{R}^2$ with corners $(1,1), (1,-1), (-1,1), (-1,-1)$ and where S_1 is the square in $\mathbb{R}^2$ with corners $(0,0), (2,0), (2,2), (0,2)$.

(a) Find an expression for the Bayes classifier and find an expression for the Bayes risk.



★ **SOLUTION:**

$$A = S_0 - (S_0 \cap S_1)$$
$$B = S_0 \cap S_1$$
$$C = S_1 - (S_0 \cap S_1)$$

$$h_*(x) = \begin{cases} 1 & x \in C \\ 0 & x \in A \\ \text{either} & x \in B \end{cases}$$

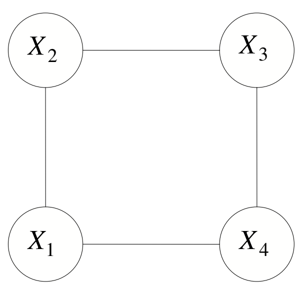
Bayes Risk $R = P(Y \neq h_*(x)) = \frac{1}{2}P(B) = \frac{1}{2}\frac{1}{4} = \frac{1}{8}$

(b) What is the best linear classifier?

Any classifier that preserves A & C. For e.g., $X_1 + X_2 = 1$

8 Graphical Models [25 points] (Jingrui)

Let $X = (X_1, X_2, X_3, X_4)$ be a random vector and consider the graph:



(a) List the local Markov properties.

★ **SOLUTION:**

$$X_1 \perp X_3 | X_2, X_4$$
$$X_2 \perp X_4 | X_1, X_3$$

(b) List the global Markov properties.

★ **SOLUTION:**

$$X_1 \perp X_3 | X_2, X_4$$
$$X_2 \perp X_4 | X_1, X_3$$

(c) Assume that all the variables are binary. Write down a graphical loglinear model for this graph.

★ **SOLUTION:**

$$\log P = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_{12} x_1 x_2 + \beta_{23} x_2 x_3 + \beta_{34} x_3 x_4 + \beta_{41} x_4 x_1$$

(d) Write down a nongraphical loglinear model for this graph.

★ **SOLUTION:** Many solutions are OK for this problem. For example,

$$\log P = \beta_0 + \beta_{12} x_1 x_2 + \beta_{23} x_2 x_3 + \beta_{34} x_3 x_4 + \beta_{41} x_4 x_1$$