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

\[ Y = m(X_1, X_2) + \epsilon \]  

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 $\beta^*$ be the best predictor, that is, $\beta^*$ minimizes $\mathbb{E}(Y - \beta X_1 X_2)^2$. Find an expression for $\beta^*$.

\[ \begin{align*}
\frac{\partial R}{\partial \beta} &= -2\mathbb{E}(Y - \beta X_1 X_2)X_1 X_2 = 0 \\
\Rightarrow \beta^* &= \frac{\mathbb{E}(Y X_1 X_2)}{\mathbb{E}(X_1^2 X_2)}
\end{align*} \]

(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 $\beta^*$ was defined in part (a).

\[ \begin{align*}
\beta^* &= \frac{\mathbb{E}(Y X_1 X_2)}{\mathbb{E}(X_1^2 X_2)} = \mathbb{E}(Y X_1 X_2) \\
&= \mathbb{E}((X_1 + X_2 + \epsilon)(X_1 X_2)) \\
&= \mathbb{E}(X_1^2 X_2 + X_1 X_2^2 + X_1 X_2) \\
&= 0
\end{align*} \]

Hence,

\[ \begin{align*}
\mathbb{E}(Y - \beta^* X_1 X_2)^2 &= \mathbb{E}(Y^2) \\
&= \mathbb{E}((X_1 + X_2 + \epsilon)^2) \\
&= \mathbb{E}(X_1^2 + X_2^2 + \epsilon^2 + 2X_1 X_2 + 2X_1 \epsilon + 2X_2 \epsilon) \\
&= 2 + \mathbb{E}(\epsilon^2)
\end{align*} \]

(c) We are given $n$ observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ from (1). Give an estimator $\hat{\beta}_n$ for $\beta^*$ and show that it is consistent.
\[ \hat{\beta} = \frac{1}{n} \sum_{i=1}^{n} X_i X_2 \xrightarrow{p} E(Y_1 X_2) \quad \frac{1}{n} \sum_{i=1}^{n} X_{1i} X_{2i}^2 \xrightarrow{p} E(X_1^2 X_2^2) \quad \therefore \hat{\beta} \to \beta \]

## 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 \( R(\theta, \hat{\theta}) \).

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

(b) Define the minimax estimator.

\[ \hat{\theta} \text{ minimizes } \sup_{\theta} R(\theta, \hat{\theta}) \]

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

\[ \hat{\theta}_\pi \text{ minimizes } R_\pi = \int R(\theta, \hat{\theta}) \pi(\theta) d\theta. \]

(d) Show that the Bayes estimator is

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

\[ R_\pi = \int \int \int (\theta - \hat{\theta})^2 f(\theta | x_1, \ldots, x_n) dx_1 \ldots dx_n. \]

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

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

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.

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

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 \(\epsilon_i\) has mean 0. Let \(X\) be the \(n \times p\) design matrix, that is, \(X(i, j) = X_{ij}\). Suppose that \(X^T X = I\) 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

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

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 \(\hat{\beta}\).

\[\textbf{★ SOLUTION:}\]

\[
\hat{\beta} = (X^T X + \lambda I)^{-1} X^T Y = (I + \lambda I)^{-1} X^T Y = \frac{1}{1+\lambda} X^T [X \beta + \epsilon] = \frac{\beta}{1+\lambda} + \frac{1}{1+\lambda} X^T \epsilon
\]

\[
\beta = E(\hat{\beta}) = \frac{\beta}{1+\lambda} \quad V(\hat{\beta}|X) = \frac{\sigma^2}{(1+\lambda)^2} X^T X = \frac{\sigma^2}{(1+\lambda)^2} I
\]

Also, \(\beta - \hat{\beta} = \frac{\lambda}{1+\lambda} \beta\)

\[
S = V(\hat{\beta}|X) = \left(\frac{\sigma^2}{(1+\lambda)^2}\right)^2 I \quad R = E(Y - X^T \hat{\beta})^2
\]

\[
E(Y - X^T \hat{\beta})^2 = E(X\beta + \epsilon - X^T \hat{\beta})^2
\]

\[
= E[(\hat{\beta} - \beta)^T X X^T (\hat{\beta} - \beta)] + \sigma^2
\]

\[
= E[(\hat{\beta} - \beta)^T X X^T (\hat{\beta} - \beta)] + 2E[(\hat{\beta} - \beta)^T X X^T (\hat{\beta} - \beta)] + E[(\hat{\beta} - \beta)^T X X^T (\hat{\beta} - \beta)] + \sigma^2
\]

\[
= \sum_{j=1}^{p} E(X_j^2) \left(\frac{\lambda}{1 + \lambda}\right)^2 \beta_j^2 + \left(\frac{\sigma}{1 + \lambda}\right)^2[2 + \left(\frac{\lambda}{1 + \lambda}\right)] + \sigma^2
\]

(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 \(Z = X^T Y\).

\[\textbf{★ SOLUTION:}\]

\[
Z = X^T Y = X^T (X \beta + \epsilon) = \beta + X^T \epsilon
\]

\[
Z \sim N(\beta, \sigma^2)
\]

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, \(P(X_i = 1) = \theta\) and \(P(X_i = 0) = 1 - \theta\) where \(0 \leq \theta \leq 1\). Consider the optimization problem

\[
\min_{\theta} f(\theta)
\]

such that \(\theta \geq 1/2\)

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

(a) What is the solution to this problem?
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.

\[ L(\theta, \lambda) = -\log \theta - 2\log(1 - \theta) + \lambda \left( \frac{1}{2} - \theta \right) \]

(c) Derive the dual problem.

\[ \frac{\partial L(\theta, \lambda)}{\partial \theta} = -\frac{1}{\theta} + \frac{2}{1 - \theta} - \lambda = 0. \text{ Therefore, } \lambda \theta^2 + (3 - \lambda)\theta - 1 = 0, \text{ and } \theta^* = \frac{\lambda - 3 + \sqrt{(\lambda - 3)^2 + 4\lambda}}{2\lambda}. \text{ The dual function: } l(\lambda) = -\log \theta^* - 2\log(1 - \theta) + \lambda \left( \frac{1}{2} - \theta^* \right). \]

(d) State the KKT conditions.

\[ -\frac{1}{\theta^*} + \frac{2}{1 - \theta^*} - \lambda^* = 0 \]
\[ \frac{1}{2} - \theta^* \leq 0 \]
\[ \lambda^* \geq 0 \]
\[ \lambda^* \left( \frac{1}{2} - \theta^* \right) = 0 \]

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 \( \mu \) that minimizes

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

where \( \lambda, \alpha > 0 \). Find \( \hat{\mu} \).

\[ \frac{\partial M}{\partial \mu} = -2(Y - \mu) + \lambda z + 2\alpha \mu \]

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

When \( \mu = 0 \), \(-2Y + \lambda z = 0 \) \( Y = \frac{\lambda}{2} \) : \( \hat{\mu} = 0 \) if \( |Y| \leq \frac{\lambda}{2} \)

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

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

\[ \hat{\mu} = \begin{cases} Y > \lambda/2 & \frac{2Y - \lambda}{2(1 + \alpha)} \\ \lambda/2 \leq |Y| \leq \lambda/2 & Y \leq \frac{\lambda}{2} \\\n0 & |Y| \leq \lambda/2 \\
\frac{2Y + \lambda}{2(1 + \alpha)} & Y < -\lambda/2 \end{cases} \]
6 Mixture Models [25 points] (Jingrui)

Let \((Z_1, Y_1), \ldots, (Z_n, Y_n)\) be generated as follows:
\[
\begin{align*}
Z_i &\sim \text{Bernoulli}(p) \\
Y_i &\sim \begin{cases} 
N(0, 1) & \text{if } Z_i = 0 \\
N(5, 1) & \text{if } Z_i = 1
\end{cases}
\end{align*}
\]

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

\[\star \text{ SOLUTION:} \]
\[f(y) = p\phi(y - 5) + (1 - p)\phi(y)\]
where \(\phi(\cdot)\) is the pdf of a standard normal distribution.

(b) Write down the likelihood function for \(p\).

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

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

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

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

\[\star \text{ SOLUTION:} \]
\[\mathbb{E}(Y) = 5p + 0(1 - p) = 5p. \text{ Let } \hat{p} = \frac{\bar{Y}}{2}. \mathbb{E}(\hat{p}) = p.\text{ According to Law of Large Numbers, } \hat{p}\text{ converges to } \mathbb{E}(\hat{p})\text{ in probability. Therefore, } \hat{p}\text{ is a consistent estimator of } p.\]

7 Classification [25 points] (Robin)

Suppose that \(P(Y = 1) = P(Y = 0) = \frac{1}{2}\) 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., \( X_1 + X_2 = 1 \)

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

★ SOLUTION:

\[
X_1 \perp X_3 | X_2, X_4 \\
X_2 \perp X_4 | X_1, X_3
\]

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

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

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