Learning theory: homework

vishvAs vAsuki

December 3, 2009

1 1

Consider $A = \left\{\left\{x \in R^p : w^T x + b \geq 0\right\}, w, b \in R^p\right\}$. This is just the set of all half spaces in R^p. Its VCD is known to be p + 1.

2 2

2.1 Problem and notation

For prior p over $\{\theta\}$, we have $r_p(d) := \int R(\theta, d)p(\theta)d\theta$.

The Bayesian estimator: $d_p := argmin_d r_p(d)$.

Assume: $r_p(d_p) = \sup_\theta R(\theta, d_p)$.

2.2 Minimax-ness of the bayesian estimator

$$
\begin{aligned}
\texttt{Let } f(d) &:= \sup_\theta R(\theta, d) \\
r_p(d) = E_\theta[R(\theta, d)] &\leq \sup_\theta R(\theta, d) = f(d) \\
\therefore r_p(d_p) = \min_d r_p(d) &\leq \min_d f(d) \\
\texttt{But } r_p(d_p) &= \sup_\theta R(\theta, d_p) = f(d_p) \\
\therefore r_p(d_p) &\leq \min_d f(d) \leq f(d_p) = r_p(d_p) \\
\therefore \min_d f(d) = \min_d \sup_\theta R(\theta, d) &= f(d_p) = r_p(d_p)
\end{aligned}
$$

Thus, d_p is also minimax.

2.2.1 Uniqueness

Suppose that d_p is the unique solution to $argmin_d r_p(d)$. Then suppose that $\exists d' :: f(d_p) = f(d') = \min_d f(d)$. Then, using earlier observations:

$$
\begin{aligned}
r_p(d') = E_\theta[R(\theta, d')] &\leq \sup_\theta R(\theta, d') \\
&= f(d') = f(d_p) = r_p(d_p) \\
&= \min_d r_p(d) \\
\therefore r_p(d') &= \min_d r_p(d)
\end{aligned}
$$

Thence, following our assumption that d_p is the unique solution to $argmin_d r_p(d)$, we conclude that $d' = d_p$. Hence, if d_p is also unique bayes, d_p is also unique minimax.

3 3

3.1 The problem setup

Observed data: $Z = (Z_1..Z_n) \in R^n$. Model: $Z_i = \theta_i + \sigma\epsilon_i$, where $\epsilon_i \sim N(0,1)$. In vector form, $Z = \theta + \sigma\epsilon$.

Loss function: $l(\hat{\theta}, \theta) = \sum(\hat{\theta}_i - \theta_i)^2 = \left\|\hat{\theta} - \theta\right\|_2^2$.

3.2 Linear shrinkage estimators

Linear shrinkage estimators: $L = \{bZ : b \in [0,1]\}$. Take $\hat{\theta} = bZ = b\theta + b\sigma\epsilon$. Then, the risk calculation is as follows:

3.2.1 Risk calculation

$$
\begin{aligned}
l(bZ, \theta) &= \left\|\hat{\theta} - \theta\right\|_2^2 \\
&= \|b\theta + b\sigma\epsilon - \theta\|_2^2 \\
&= \|(b-1)\theta + b\sigma\epsilon\|_2^2 \\
&= (b-1)^2\theta^T\theta + (b\sigma)^2\epsilon^T\epsilon + 2(b-1)b\sigma\theta^T\epsilon \\
R(bZ, \theta) &= E_Z[l(bZ, \theta)] = E_\epsilon[l(bZ, \theta)] \\
&= (b-1)^2\theta^T\theta + (b\sigma)^2 E_\epsilon[\epsilon^T\epsilon] + E_\epsilon[2(b-1)b\sigma\theta^T\epsilon] \\
&= (b-1)^2\theta^T\theta + (b\sigma)^2 n E_\epsilon[\epsilon_i^2] \texttt{ As last term is 0, } \epsilon_i \texttt{ are iid.} \\
&= (b-1)^2\theta^T\theta + (b\sigma)^2 n
\end{aligned}
$$

3.2.2 Minimax estimator for unrestricted parameters

Let $T = R^p$. $\sup_{\{\theta \in T\}} R(bZ, \theta) = \infty$ if $b \neq 1$ and $(b\sigma)^2 n$ otherwise. So, Z is the minimax estimator. $R(Z, \theta) = \sigma^2 n$.

3.2.3 When parameters restricted to a ball

Let $T = \left\{\theta : \|\theta\|_2^2 \leq R^2\right\}$. Then:

$$
\begin{aligned}
f(b) = \sup_{\{\theta \in T\}} R(bZ, \theta) &= (b-1)^2 R^2 + (b\sigma)^2 n \\
f(\hat{b})' &= 2(\hat{b}-1)R^2 + 2\hat{b}\sigma^2 n = 0 \texttt{ Min wrt b} \\
\hat{b} &= (R^2 + \sigma^2 n)^{-1} R^2 \\
R(\hat{b}Z, \theta) &= (\hat{b}-1)^2 \theta^T \theta + (\hat{b}\sigma)^2 n
\end{aligned}
$$

Comparison with Z and admissibility We have seen that $R(Z, \theta) = \sigma^2 n$. Does this imply inadmissibility of $\hat{b}Z$? This implication is not true if
if $\exists \theta : [R(\hat{b}Z, \theta) < R(Z, \theta)]$ is true. To see that this condition holds, consider any $\theta : \|\theta\|_2^2 < \sigma^2 n$.

Even though we still have not proved the admissibility of $\hat{b}Z$, we see that comparizon with Z does not let us conclude that $\hat{b}Z$ is inadmissible.

3.2.4 When parameters restricted to hyper-ellipse

Let 1 be the vector 1^n, and e_i the the ith column of I_n.

Let $T = \left\{\theta : \sum_j a_j^2 \theta_j^2 \leq c^2\right\} = \left\{\theta : \theta^T A \theta \leq c^2 1, A = diag(a_1^2..a_n^2)\right\}$. Consider $L = \{WZ : W = diag(w_1, ..w_n)\}$. Then:

$$
\begin{aligned}
WZ &= W\theta + \sigma W \epsilon \\
\|WZ - \theta\|_2^2 &= \|(W-I)\theta + \sigma W \epsilon\|_2^2 \\
&= \theta^T (W-I)^2 \theta + \sigma^2 \epsilon^T W^2 \epsilon + 2\sigma \epsilon^T W^T (W-I)\theta \\
R(WZ, \theta) &= E_Z[\|WZ - \theta\|_2^2] = E_\epsilon[\|WZ - \theta\|_2^2] \\
&= \theta^T (W-I)^2 \theta + E_\epsilon[\sigma^2 \epsilon^T W^2 \epsilon] \\
t(W) &:= argsup_{\theta \in T} \theta^T (W-I)^2 \theta \\
&= (\frac{c}{a_i}) e_i : i = argmax_j (w_j - 1)^2 (\frac{c}{a_j})^2 \\
\sup_{\theta \in T} R(WZ, \theta) &= t(W)^T (W-I)^2 t(W) + \sigma^2 1^T W^2 1 \\
\min_w \sup_{\theta \in T} R(WZ, \theta) &= ?
\end{aligned}
$$

[**Incomplete**]

4 4

4.1 Notation

B_1: l_1 unit ball in R^p. Want to find lower bound of $M(\epsilon, B_1, \|.\|_2)$.

4.2 The set S

$m \in N, S = \{u \in \{-1,0,1\}^p : \|u\|_1 = 2m\}$. Its cardinality $\#S = \binom{p}{2m}2^{2m}$, considering $\binom{p}{2m}$ ways of picking positions for non 0 elements, and 2^{2m} ways of assigning values from $\{1,-1\}$ to those positions.

4.2.1 Number of vectors with low hamming distance from v

h(u, v) is the hamming distance on S.

Fix $v \in S$. Consider
$close(v) = \{u \in S : h(u,v) \leq m\} \subseteq \{u \in \{-1,0,1\}^p : h(u,v) \leq m\}$. The cardinality of the latter set is bounded above by $\binom{p}{m}3^m$, where $\binom{p}{m}$ counts the number of ways of choosing the positions at which u potentially differs from v, and 3^m counts the possible values u has in those positions.

Thus, $\#close(v) \leq \binom{p}{m}3^m$.

4.2.2 Number of vectors with low hamming distance from A

Take any $A \subseteq S$ with cardinality $a_m = \binom{p}{2m}/\binom{p}{m}$. Then, using the previous result,
$\#\{u \in S : \exists v \in A, h(u,v) \leq m\} \leq |A|\#close(v) \leq \binom{p}{2m}3^m < \binom{p}{2m}2^{2m} = \#S$.

Existance of v in S atleast m-away from A From this, we see that there $\exists y \in S - A : \forall v \in A : h(v,y) > m$. This also implies that, $\forall A \subseteq S : |A| \leq a_m$, $\exists y \in S - A : \forall v \in A : h(v,y) > m$.

4.2.3 Packing set for S

The following process constructs a packing set for S. Start with $A = \{\}$. Do the following while $|A| \leq a_n$: find $y \in S - A : \forall v \in A : h(v,y) > m$, set $A = A \cup \{y\}$. We are sure that there always exists such a y as long as $|A| \leq a_n$ due to a previous result.

Using the above process, we have constructed a packing set A for S, whose elements are at least m apart in the h metric.

4.2.4 Relating hamming distance to the sq euclidian norm

Take $\forall u, v \in A$. $|u-v|$ has either 1 or 2 in atleast m positions. So, $\|u-v\|_2^2 > m$, and $\forall u, v \in A : \|u-v\|_2 > \sqrt{m}$.

4.3 A packing set for the unit ball

Consider the set S and its packing A described above. Take $A_1 = \left\{(2m)^{-1}v : v \in A\right\}$. Every v in A has exactly 2m positions with ±1 values, and we had: $\forall v \in A, \|v\|_1 = 2m$. So, $\forall v \in A_1 : \|v\|_1 = 1$, and $A_1 \subseteq B_1$.

As $\forall u, v \in A : \|u - v\|_2 > \sqrt{m}$, we have $\forall u, v \in A_1 : \|u - v\|_2 > (2m)^{-1}\sqrt{m}$.

So, for any $\epsilon \leq 1/(2\sqrt{m})$ or $m \leq (2\epsilon)^{-2}$, we have $a_m \leq M(\epsilon, B_1, \|.\|_2)$.

4.3.1 A rough lower bound on a_m

$a_m = \binom{p}{2m}/\binom{p}{m} \geq (\frac{p}{2m})^{2m}/(\frac{p^m}{m!}) \geq \frac{p^m k^m}{m^m}$ for a constant k. In the first inequality, we apply lower and upper bounds $(\frac{a}{b})^b < \binom{a}{b} < a^b/b!$. In the second inequality, we use an inequality from Stirling's approximation for m!, whose use is justified because m tends to be large.