Stat 928: Statistical Learning Theory **Lecture: 8**

Feature Selection in the Non-Orthogonal Case

Instructor: Sham Kakade

1 Feature Selection

Our goal now is to understand how to select the best q features out of p possible features. Throughout this analysis, let us assume that:

$$Y = X\beta + \eta,$$

where $Y \in R^n$ and $X \in \mathbb{R}^{n\times p}$. We assume that the support of β is q.

1.1 Empirical Risk Minimization

Recall that:

$$L(w) = \frac{1}{n}\mathbb{E}\|Xw - Y\|^2 = \frac{1}{n}\|Xw - \mathbb{E}[Y]\|^2 + \sigma^2$$

Define our "empirical loss" as:

$$\hat{L}(w) = \frac{1}{n}\|Xw - Y\|^2$$

which has no expectation over Y. Note that for a fixed w

$$\mathbb{E}[\hat{L}(w)] = L(w)$$

e.g. the empirical loss is an unbiased estimate of the true loss.

Suppose we knew the support size q. One algorithm is to simply find the estimator which minimizes the empirical loss and has support only on q coordinates.

In particular,

$$\hat{\beta}_q = \inf_{\text{support}(w)\leq q} \hat{L}(w)$$

where the $\inf$ is over vectors with support size q.

We are interested in, with probability,

$$L(\hat{\beta}_q) - L(\beta) \leq ??$$

Recall the risk is:

$$\mathbb{E}_Y[L(\hat{\beta}_q)] - L(\beta) \leq ??$$

where the expectation is over Y.

2 How accurate are the true and empirical losses?

Let's ignore the feature selection issue for a moment and just return to linear regression, and consider the case where it may be that $\mathbb{E}[Y] \neq \beta X$, e.g. let's not assume that model is correct. This will be relevant since we consider subspaces which may not be the best subspace.

Lemma 2.1. *Let β be the best linear predictor (i.e. it may be that $\mathbb{E}[Y] \neq \beta X$, but β is still the best linear predictor.) Let $\hat{\beta}$ be the least squares estimate. We have that:*

$$L(\hat{\beta}) - L(\beta) = \frac{1}{n}\|\Pi\eta\|^2$$

We also have that:

$$\hat{L}(\beta) - \hat{L}(\hat{\beta}) = \frac{1}{n}\|\Pi\eta\|^2$$

Proof. Let $\hat{Y}$ be our prediction of $E[Y]$, i.e.:

$$\hat{Y} = \Pi Y = X\hat{\beta}$$

Note that:

$$L(\hat{\beta}) - L(\beta) = \frac{1}{n}\|\Pi\mathbb{E}[Y] - \Pi Y\|^2 = \frac{1}{n}\|\Pi\eta\|^2$$

(we also saw this in Lecture 2, lemma 3.2).

Now note that for all w,

$$\hat{L}(w) = \|Xw - Y\|^2 = \|Xw - \Pi Y + (Y - \Pi Y)\|^2 = \hat{L}(\hat{\beta}) + \|Xw - \Pi Y\|^2$$

where the cross term is 0 due to that $\hat{\beta}$ is the best linear predictor.

Hence, using $w = \beta$,

$$\hat{L}(\beta) - \hat{L}(\hat{\beta}) = \frac{1}{n}\|\Pi\mathbb{E}[Y] - \Pi Y\|^2 = \frac{1}{n}\|\Pi\eta\|^2$$

which completes the proof. □

2.1 Comment: Accuracy of the empirical loss

But what about:

$$L(\hat{\beta}) - \hat{L}(\hat{\beta}) = ??$$

and

$$L(\beta) - \hat{L}(\beta) = ??$$

It turns out that (with high probability) these are not all that small (they are $O(\sqrt{1/n})$ (ignoring dimension dependencies).

Assume that η has variance σ^2 in each coordinate. For this case, note that the empirical loss is just sum of η_i^2, since $Y = X\beta + \eta$

Note that we can write:

$$L(\beta) - \hat{L}(\beta) = \frac{1}{n}\sum_i(\sigma^2 - \eta_i^2)$$

By the central limit theorem, we know that for large n

$$\frac{1}{n}\sum_i(\sigma^2 - \eta_i^2) \approx 1/\sqrt{n}$$

Hence:

$$L(\beta) - \hat{L}(\beta) \approx 1/\sqrt{n}$$

Hence, we expected B (in the empirical process) to be $1/\sqrt{n}$.

3 Understanding Feature Selection

A key question is how does the loss of any least squares estimate on $\mathcal{S}$ related to the loss of β?

Lemma 3.1. *For any $\mathcal{S}$,*

$$L(\beta_{\mathcal{S}}) - L(\beta) = \hat{L}(\beta_{\mathcal{S}}) - \hat{L}(\beta) - \frac{1}{n}(X\beta_{\mathcal{S}} - X\beta) \cdot \eta$$

where $\hat{\beta}_{\mathcal{S}}$ is the least squares estimate on $\mathcal{S}$ and β is the best linear predictor.

Proof. Observe

$$\begin{aligned}
\hat{L}(\beta_{\mathcal{S}}) &= \frac{1}{n}\|X\beta_{\mathcal{S}} - Y\|^2 \\
&= \frac{1}{n}\|X\beta_{\mathcal{S}} - (X\beta + \eta)\|^2 \\
&= L(\beta_{\mathcal{S}}) - L(\beta) + \frac{1}{n}(X\beta_{\mathcal{S}} - X\beta) \cdot \eta + \frac{1}{n}\|\eta\|^2 \\
&= L(\beta_{\mathcal{S}}) - L(\beta) + \frac{1}{n}(X\beta_{\mathcal{S}} - X\beta) \cdot \eta + \hat{L}(\beta)
\end{aligned}$$

which completes the proof. □

3.1 Feature Selection Analysis

Lemma 3.2. *Let the ERM subspace $\hat{\mathcal{S}}$ be such that have:*

$$\hat{L}(\hat{\beta}_{\hat{\mathcal{S}}}) - \hat{L}(\beta) \leq 0$$

We ahve

$$L(\beta_{\hat{\mathcal{S}}}) - L(\beta) \leq -\frac{1}{n}(X\beta_{\hat{\mathcal{S}}} - X\beta) \cdot \eta + \frac{1}{n}\|\Pi_{\hat{\mathcal{S}}}\eta\|^2$$

where $\beta_{\hat{\mathcal{S}}}$ is best linear predictor on this subspace.

Proof. Use that $\hat{L}(\hat{\beta}_{\hat{\mathcal{S}}})$ is close to $\hat{L}(\beta_{\hat{\mathcal{S}}})$ by $\frac{1}{n}\|\Pi_{\hat{\mathcal{S}}}\eta\|^2$. □

Hence we must bound the last two terms for the ERM subspace. Instead, we will consider bounding the following for all $\mathcal{S}$ (as this implies a bound on the ERM subspace)

$$\frac{1}{n}(X\beta_{\mathcal{S}} - X\beta) \cdot \eta \leq ??$$

and

$$\frac{1}{n}\|\Pi_{\mathcal{S}}\eta\|^2 \leq ??$$

Lemma 3.3. *We have that:*
$$Var(\frac{1}{n}(X\beta_{\mathcal{S}} - X\beta) \cdot \eta) = \frac{1}{n}(L(\beta_{\mathcal{S}}) - L(\beta))$$

For the first term, we have that:
$$\frac{1}{n}(X\beta_{\mathcal{S}} - X\beta) \sim N(0, \frac{1}{n}(L(\beta_{\mathcal{S}}) - L(\beta)))$$

Hence for any given $\mathcal{S}$, we have that:
$$|\frac{1}{n}(X\beta_{\mathcal{S}} - X\beta)| \leq \sqrt{\frac{2(L(\beta_{\mathcal{S}}) - L(\beta))\log(2/\delta)}{n}} \leq \frac{1}{2}(L(\beta_{\mathcal{S}}) - L(\beta)) + O(\frac{\log(1/\delta)}{n})$$
using $2ab \leq a^2 + b^2$, which implies (with an $a = \sqrt{(L(\beta_{\mathcal{S}}) - L(\beta))/2}$).

Now using the χ^2 tail bound, we have that:
$$\|\Pi_{\mathcal{S}}\eta\|^2 \leq q + 2\sqrt{q\ln(1/\delta)} + 2q\ln(1/\delta) \leq O(q + \ln(1/\delta))$$

Hence we have that:

Theorem 3.4. *We have that with probability greater than $1-\delta$, for the ERM $\hat{\beta}_q$ (constrained to only choose q features):*
$$L(\hat{\beta}_q) - L(\beta) \leq O\left(\frac{q + \log(\binom{q}{p}/\delta)}{n}\right)$$

4 χ^2 Tail Bound

Let $X_i \sim N(0,1)$ be independent Gaussians, then the distribution of $Z = \sum_{i=1}^n X_i^2$ is χ^2 with n degrees of freedom.

This variable is important for analyzing least squares regression.

Theorem 4.1. *Let $X_i \sim N(0,1)$ be independent Gaussians, then the distribution of $Z = \sum_{i=1}^n X_i^2$ is χ^2. We have that (for the upper tail):*
$$P(Z/n \geq 1 + \epsilon) \leq \exp\left[-\frac{n}{2}(\epsilon - \log(1+\epsilon))\right]$$
One useful upper bound (for obtaining sharp constants) is:
$$\exp\left[-\frac{n}{2}(\epsilon - \log(1+\epsilon))\right] \leq \exp\left[-\frac{n}{2}(1 + \epsilon - \sqrt{1+2\epsilon})\right]$$
A bound that is more comparable to the Bennet-style bound is:
$$\exp\left[-\frac{n}{2}(\epsilon - \log(1+\epsilon))\right] \leq \exp[-n\epsilon^2/(4+4\epsilon)].$$
(note the difference between the upper and lower tail).

For the lower tail:
$$P(Z/n \leq 1 - \epsilon) \leq \exp[-n\epsilon^2/4].$$

Hence, with probability $1-\delta$:
$$Z/n \leq 1 + 2\sqrt{\ln(1/\delta)/n} + 2\frac{\ln(1/\delta)}{n}$$
and with probability $1-\delta$:
$$Z/n \geq 1 - 2\sqrt{\ln(1/\delta)/n}.$$

The logarithmic moment generating function of X_i^2 for $\lambda < 0.5$ is

$$\Gamma(\lambda) = \ln E e^{\lambda X_i^2} = -0.5 \ln(1 - 2\lambda),$$

and $EX_i^2 = 1$.

Proof. We only prove the upper tail. The lower tail is simpler to prove in that we can use the bound $log(1 + x) > 1 + x - x^2/2$ for $x > 0$.

From the moment method, we must constrain $\lambda < -.5$, or, equivalently, set $\Gamma(\lambda) = \infty$ for $\lambda \geq 0.5$. Hence,

$$I(1 + \epsilon) = \inf_{0.5 > \lambda > 0} [-\lambda(1 + \epsilon) - 0.5 \ln(1 - 2\lambda)] = -\frac{1}{2} (\epsilon - \log(1 + \epsilon))$$

where the inf is achieved at $1 + \epsilon = \frac{1}{1-2\lambda}$ or equivalently $\lambda = \frac{\epsilon}{2(1+\epsilon)}$.

The first claim is completed by noting that $\log(1 + \epsilon) \leq \sqrt{1 + 2\epsilon} - 1$, for $\epsilon > 0$. To see this, first note equality at $\epsilon = 0$. Also, note that derivative on the left hand side is:

$$\frac{1}{1 + \epsilon} =\leq \frac{1}{\sqrt{1 + 2\epsilon}}$$

where the right hand side is the derivative of $\sqrt{1 + 2\epsilon}$.

For the second claim, the proof is completed by noting that the function $f(x) = (x - \log(1 + x)) * (1 + x)$. Note that $f'(x) = 2x - \log(1 + x)$, $f''(x) = (1 + 2x)/(1 + x)$, and $f'''(x) = 1/(1 + x)^2 >= 0$. So $f(x) >= x^2/2$.

The rest of the proof is straight forward. $\square$