CSE 546: Machine Learning **Lecture 5 extras**

Optional Reading: Feature Selection Risk

Instructor: Sham Kakade

1 Comments

The following is a proof of the risk bound for feature selection that we considered in class (we will actually provide a slightly stronger bound in that we provide a bound that holds with high probability, rather than just in expectation).

This proof is a little more involved than others we have seen. Later in the class, we will see a simpler argument which qualitatively shows why the dependence on the dimension d should be logarithmic.

2 Feature Selection

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

$$\mathbf{Y} = \mathbf{X}w_* + \eta,$$

where $\eta \sim \mathcal{N}(0, \sigma^2)$, $\mathbf{Y} \in R^n$ and $\mathbf{X} \in \mathbb{R}^{n\times d}$. We assume that the support of w_* is s.

2.1 Subset selection

Note that:

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

Define our "empirical loss" as:

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

which has no expectation over $\mathbf{Y}$. 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 s. One algorithm is to simply find the estimator which minimizes the empirical loss and has support only on s coordinates.

In particular,

$$\hat{w}_s = \inf_{\text{support}(w)\leq s} \hat{L}(w)$$

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

We will bound the following quantity:

$$L(\hat{w}_s) - L(w_*) \leq ??$$

(In particular, we will provide a bound that holds with high probability.) Recall the risk is:

$$\mathbb{E}_{\mathbf{Y}}[L(\hat{w}_s)] - L(w_*) \leq ??$$

where the expectation is over $\mathbf{Y}$.

The main theorem is as follows:

Theorem 2.1. *(a high probability bound) We have that with probability greater than $1-\delta$,*

$$L(\hat{w}_s) - L(w_*) \leq c\frac{(s + \log(\binom{d}{s}/\delta))}{n}\sigma^2 \leq c\frac{(s + s\log(d/\delta))}{n}\sigma^2$$

where $\binom{d}{s}$ is the number of subsets of size s and c is a universal constant.

3 How accurate are the true and empirical losses?

Let us ignore the feature selection issue for a moment and just return to linear regression. It will be important for us to consider the case where it may be that $\mathbb{E}[\mathbf{Y}] \neq \mathbf{X}w$, e.g. we need to consider the case where the model is not correct. This is relevant as we will consider least squares estimates on subsets which may not be the best subset.

Lemma 3.1. *Let w_* be a minimizer of $L(w)$ (where it may be the case that $\mathbb{E}[\mathbf{Y}] \neq \mathbf{X}w_*$). Let $\Pi = UU^\top$ where U is $n \times d$ left orthogonal matrix of the thin SVD of $\mathbf{X}$, so Π is a projection matrix. Let $\hat{w}$ be the least squares estimate. We have that:*

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

We also have that:

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

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

$$\hat{\mathbf{Y}} = \Pi\mathbf{Y} = \mathbf{X}\hat{w}$$

Note that:

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

(we also saw this in Lecture 2).

Now note that for all w,

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

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

Hence, using $\mathbf{X}w_* = \Pi\mathbb{E}[\mathbf{Y}]$,

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

which completes the proof. □

4 Feature Selection Analysis

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

Lemma 4.1. *For any subset $\mathcal{S}$,*

$$L(w_{\mathcal{S}}) - L(w_*) = \hat{L}(w_{\mathcal{S}}) - \hat{L}(w_*) - \frac{1}{n}(\mathbf{X}w_{\mathcal{S}} - \mathbf{X}w_*) \cdot \eta$$

where $w_{\mathcal{S}}$ is the best fit line on $\mathcal{S}$ and w_* is the best linear predictor overall.

Proof. Observe

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

which completes the proof. □

The following lemma is immediate:

Lemma 4.2. *Let the selected subset $\hat{\mathcal{S}}$ be such that:*

$$\hat{L}(\hat{w}_{\hat{\mathcal{S}}}) - \hat{L}(w_*) \leq 0$$

(i.e. our selected subset will have empirical loss that is smaller than w_*). We have

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

where $w_{\hat{\mathcal{S}}}$ is best linear predictor on this subset.

Proof. Use the previous lemma and that $\hat{L}(\hat{w}_{\hat{\mathcal{S}}}) - \hat{L}(w_{\hat{\mathcal{S}}}) = \frac{1}{n}\|\Pi_{\hat{\mathcal{S}}}\eta\|^2$. □

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

$$\frac{1}{n}(\mathbf{X}w_{\mathcal{S}} - \mathbf{X}w_*) \cdot \eta \leq ??$$

and

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

Lemma 4.3. *We have that:*

$$Var(\frac{1}{n}(\mathbf{X}w_{\mathcal{S}} - \mathbf{X}w_*) \cdot \eta) = \frac{1}{n}(L(w_{\mathcal{S}}) - L(w_*))$$

We are now ready to complete the proof of the main theorem. For the first term, we have that:

$$\frac{1}{n}(\mathbf{X}w_{\mathcal{S}} - \mathbf{X}w_*) \sim \mathcal{N}(0, \frac{1}{n}(L(w_{\mathcal{S}}) - L(w_*)))$$

Hence, using the Gaussian tail bound (see the notes on large deviations), for any given $\mathcal{S}$, we have that:

$$|\frac{1}{n}(\mathbf{X}w_{\mathcal{S}} - \mathbf{X}w_*)| \leq \sqrt{\frac{2(L(w_{\mathcal{S}}) - L(w_*))\log(2/\delta)}{n}} \leq \frac{1}{2}(L(w_{\mathcal{S}}) - L(w_*)) + 4(\frac{\log(2/\delta)}{n})$$

using $2ab \leq a^2 + b^2$ (with $a = \sqrt{(L(w_{\mathcal{S}}) - L(w_*))/2}$).

Now using the χ^2 tail bound (see the notes on large deviations), we have that:

$$\|\Pi_{\mathcal{S}}\eta\|^2 \leq \left(s + 2\sqrt{s\ln(1/\delta)} + 2\ln(1/\delta)\right)\sigma^2 \leq 4(s + \ln(1/\delta))\sigma^2$$

Note that we desire both of these bounds to hold on the selected subset $\hat{\mathcal{S}}$. To do this, we actually will demand that the bounds hold for *all* subsets $\mathcal{S}$. In doing so, if we replace δ with $\delta/\binom{d}{s}$, then the previous bounds hold for all subsets (this is the union bound). This completes the proof.