Stat 928: Statistical Learning Theory **Lecture: 7**

Feature Selection, Empirical Risk Minimization, and The Orthogonal Case

Instructor: Sham Kakade

1 A Few Relevant Features

Our goal now is to understand how to select the best q features our 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:

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

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.

1.2 Coordinate dependence?

Clearly, the coordinates system is important here, as the "support" is defined with respect to this coordinate system. However, note that the 'scale' in each coordinate is irrelevant here. In particular, note the empirical risk minimization does not depend on the scale of each coordinate.

1.3 Subset Estimation

Let $\mathcal{S}$ denote a subset of $[q]$ (here $[q]$ denotes the set of features $\{1, 2, \dots q\}$). Let us specify the estimator on the subset $\mathcal{S}$ (e.g. the subspace due to the coordinates $\mathcal{S}$).

Denote the restriction of X to the subset $\mathcal{S}$ to be $X_{\mathcal{S}} \in \mathbb{R}^{n \times q}$. We have that:

$$\beta_S = \frac{1}{n} X_{\mathcal{S}}^{\dagger} Y$$

where † denotes the pseudo inverse.

$$\beta_q = \arg\min_{\mathcal{S}} \|X_{\mathcal{S}} \beta_{\mathcal{S}} - Y\|^2$$

1.4 Computational Issues?

Note there are $\binom{p}{q}$ subsets of size q. Naively, this optimization would involve searching these subsets. What are cases under which this optimization can be done efficiently?

More generally, are there special cases where can find an estimator which has low risk in this setting?

{**SK:** *example*}

2 The Orthogonal Case

Note that the MLE over *all* coordinates is:

$$\hat{\beta} = \frac{1}{n}(X^\top X)^{-1} X^\top Y = \frac{1}{n} X^{\dagger} Y$$

by assumption.

Let us suppose that our design matrix is orthogonal. In particular, suppose that:

$$\Sigma = \frac{1}{n} X^\top X = \text{diagonal}$$

Under the diagonal assumption, without loss of generality, we can assume without loss of generality that:

$$\Sigma = \mathrm{I}$$

(by rescaling each coordinate).

Here we have that j-th coordinate of the (global) MLE β_j is just correlation between j-th dimension and Y.

$$\hat{\beta}_j = \frac{1}{n}(X^\top Y)_j = \frac{1}{n} X_j \cdot Y = \frac{1}{n} \sum_i X_{i,j} Y_i := \hat{\mathbb{E}}[x_j y]$$

where X_j is the j-th column of X.

Also, by the identity assumption, we have that j-th coordinate of the MLE on $\mathcal{S}$ is just the restriction of

$$[\hat{\beta}_{\mathcal{S}}]_j = \hat{\beta}_j = \frac{1}{n}\sum_i X_{i,j} Y_i$$

Hence, estimation $\hat{\beta}_{\mathcal{S}}$ for all $\mathcal{S}$ is easy in this case.

2.1 Regret in the Orthogonal Case

Let $\mathcal{S}^\star$ be the optimal support set (e.g. the support set of the true β).

We can write the regret as:

$$L(\hat{\beta}_{\mathcal{S}}) - L(\beta) = \frac{1}{n}\|X\hat{\beta}_{\mathcal{S}} - \mathbb{E}[Y]\|^2 = \frac{1}{n}\|X\hat{\beta}_{\mathcal{S}} - X\beta\|^2 = \|\hat{\beta}_{\mathcal{S}} - \beta\|^2 = \sum_{j\notin\mathcal{S}} \beta_j^2 + \sum_{j\in\mathcal{S}} (\beta_j - [\hat{\beta}_S]_j)^2$$

2.2 High Probability Bounds on the Regret

Suppose that η is sub-Gaussian with constant σ.

Hence, with probability greater than $1-\delta$, that:

$$\hat{\beta}_j \leq \beta_j + \sqrt{\frac{2\sigma^2 \log(1/\delta)}{n}} \tag{1}$$

and also that:

$$\hat{\beta}_j \geq \beta_j - \sqrt{\frac{2\sigma^2 \log(1/\delta)}{n}}$$

(using our sub-Gaussian tail bound, from the previous lecture).

Note that union bounds states that (for events $\mathcal{E}_1$ to $\mathcal{E}_k$) that:

$$\Pr(\mathcal{E}_1 \text{ or } \mathcal{E}_2 \ldots \text{ or } \mathcal{E}_k) \leq \sum_j \Pr(\mathcal{E}_j)$$

(question: when is this sharp?).

Now consider the following $2p$ events: one of the above 2 equations fail for some coordinate j. Note that if use $\delta/2p$ in the above the cumulative failure probability is less than:

$$\Pr(\text{ any failure }) \leq \sum_j \Pr(\text{ one-sided failure for j}) \leq 2p(\delta/2p) = \delta$$

Hence, we have that, with probability greater than $1-\delta$, that:

$$\max_j |\hat{\beta}_j - \beta_j| \leq \sqrt{\frac{2\sigma^2 \log(2p/\delta)}{n}}$$

so that δ is a failure probability with respect to any event.

The following theorem is immediate for the empirical risk minimization:

Lemma 2.1. *(Regret Bound) Let $\hat{\beta}_{\mathcal{S}}$ be the empirical risk minimizer over all subsets of size no more than q. We have that with probability greater than $1-\delta$:*

$$L(\hat{\beta}_{\mathcal{S}}) - L(\beta) \leq \frac{4q\sigma^2 \log(2p/\delta)}{n}$$

Proof. For those features $j \in \mathcal{S}^\star$, we have that:

$$\sum_{j \in \mathcal{S}^\star} (\hat{\beta}_j - \beta_j)^2 \leq \frac{2q\sigma^2 \log(2p/\delta)}{n}$$

Also, by construction, there are on q features $j \notin \mathcal{S}^\star$ which are non-zero. Hence,

$$\sum_{j \notin \mathcal{S}^\star} (\hat{\beta}_j - \beta_j)^2 \leq \frac{2q\sigma^2 \log(2p/\delta)}{n}$$

which completes the proof. □

2.3 A Better Constant with Variable Selection

Note that since:

$$\max_j |\hat{\beta}_j - \beta_j| \leq \sqrt{\frac{2\sigma^2 \log(2p/\delta)}{n}}$$

we are confident that β_j is non-zero if

$$|\hat{\beta}_j| \geq \sqrt{\frac{2\sigma^2 \log(2p/\delta)}{n}} \tag{2}$$

So a different algorithm is to use only those $\hat{\beta}_j$ for which the above condition holds — else we just set $\hat{\beta}_j = 0$. Note this algorithm doesn't need to know q.

Lemma 2.2. *If we use the estimator defined above (e.g. use only those j for which Equation 2 holds), then, with probability greater than $1-\delta$*

$$L(\hat{\beta}) - L(\beta) \leq \frac{2\sigma^2 \log(2p/\delta)}{n}$$

Proof. Note that with probability greater than $1-\delta$ we do not include an incorrect coordinate. For the $j \in \mathcal{S}^\star$, the error contribution for each coordinate is:

$$|\hat{\beta}_j - \beta_j|^2 \leq \frac{2\sigma^2 \log(2p/\delta)}{n}$$

The proof is completed by summing over these q errors. □

2.4 What about the risk?

The above provides high probability bounds. What can we say about the risk? In other words, what is:

$$\mathbb{E}[\|\hat{\beta} - \beta\|_\Sigma^2 = \mathbb{E}[L(\hat{\beta})] - L(\beta) = ??$$

where the expectation is with respect to sample Y.

Note that for the ERM algorithm, it only include q features (at most). Also note that:

$$L(\hat{\beta})] - L(\beta) = \sum_{j \notin \mathcal{S}} \beta_j^2 + \sum_{j \in \mathcal{S}} (\beta_j - [\hat{\beta}_{\mathcal{S}}]_j)^2 \leq 2q \max_j |\hat{\beta}_j - \beta_j|$$

Hence:

$$\mathbb{E}[\|\hat{\beta} - \beta\|_\Sigma^2 \leq 2q\mathbb{E}[\max_j |\hat{\beta}_j - \beta_j|]$$

We will return to this next lecture.

3 The Non-Orthogonal Case

Note now there is no reason that an estimate the ERM can be computed easily (and there are hardness results to this effect). However, for now, let us ignore this issue and examine the performance of the "subset selection" ERM algorithm.

Let $\hat{Y}_{\mathcal{S}}$ be our prediction of $E[Y]$ using our estimate on $\mathcal{S}$, i.e.

$$\hat{Y}_{\mathcal{S}} = X\beta_{\mathcal{S}} = \Pi_{\mathcal{S}} Y$$

Also, note the best linear predictor using only those features in $\mathcal{S}$, is:

$$\Pi_{\mathcal{S}} \mathbb{E}[Y]$$

where $\Pi_{\mathcal{S}}$ is the projection of Y onto the subspace spanned by X_j for $j \in \mathcal{S}$.

3.1 Naively, an empirical process theory approach

Let's attempt a simple (and too loose) attempt at a proof.

Let B be the maximal deviation:

$$B = \max_{\mathcal{S}} |\hat{L}(\hat{\beta}_{\mathcal{S}}) - L(\hat{\beta}_{\mathcal{S}})|$$

for some constant B. Later we will see that B is an "empirical process".

For the ERM β_q, say which uses $\mathcal{S}$, this would imply that;

$$L(\hat{\beta}_{\mathcal{S}}) \leq \hat{L}(\hat{\beta}_{\mathcal{S}}) + B \leq \hat{L}(\hat{\beta}_{\mathcal{S}^\star}) + BL(\hat{\beta}_{\mathcal{S}^\star}) + 2B$$

Furthermore, at least in expectation, we have previously shown that:

$$\mathbb{E}[L(\hat{\beta}_{\mathcal{S}^\star}) - L(\beta_{\mathcal{S}^\star})] = \frac{q\sigma^2}{n}$$

Thus we have that:

$$\mathbb{E}[L(\hat{\beta}_{\mathcal{S}}) - L(\hat{\beta})] \leq \frac{q\sigma^2}{n} + 2\mathbb{E}[B]$$

where $\beta_{\mathcal{S}^\star} = \beta$.

The issue then is that:

$$\mathbb{E}[B] \leq ??$$

It turns out that this approach does not lead to sharp analysis (as $B \approx 1/\sqrt{n}$).

4 How accurate are the 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 4.1. *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:*

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

Alternatively, we have that the difference in losses is:

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

and, in expectation, this difference is well behaved, i.e.

$$\mathbb{E}[L(\hat{\beta}) - \hat{L}(\hat{\beta}) - (L(\beta) - \hat{L}(\beta))] = \frac{2p\sigma^2}{n}$$

where p is the dimension of column space of X.

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. □

4.1 Comment: Accuracy of the empirical loss

Let's consider the simplest case where $\mathcal{S} = \mathcal{S}^\star$, e.g. the best subset. So we are using the true β.

Clearly,

$$L(\beta) - \hat{L}(\beta) = 0$$

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 $1/\sqrt{n}$.

4.2 The quadratic form

The key quantity of interest is:

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

If η is a Gaussian, then note that $\Pi_{\mathcal{S}}\eta$ is also a Gaussian vector of length q, i.e.

$$\Pi_{\mathcal{S}}\eta\sim N(0,\mathrm{I}_{q\times q})$$

Hence, in distribution,

$$\frac{1}{n}\|\Pi_{\mathcal{S}}\eta\frac{\sigma^2}{n}\|^2\frac{\text{in distribution}}{=}\sum_i x_i^2$$

where x_i are standard normal distributions.

Note that the variance of the right hand side is $k\sigma^2/n$, we expect that:

$$\frac{1}{n}\|\Pi_{\mathcal{S}}\eta\frac{\sigma^2}{n}\|^2\approx\frac{q\sigma^2}{n}+\sqrt{\frac{q\sigma^2}{n}}+??$$

Let us make this precise.

Comment: Here we have utilized a crucial property of the Gaussian noise. That a Gaussian is spherically symmetric (so a rotation preserve independence). The proof for sub-Gaussian η is much trickier.