Statistical Machine Learning Lecture 6— Persistence Sept. 25, 2014

Lecturer: Prof. Homrighausen Scribe: H. R. Scharf

0 A short discussion on Safe Rules - Chuck Vollmer

When dealing with data with large n and p, computation of the LASSO estimate via traditional methods (eg: `glmnet` or `lars` packages in `R`) can become burdensome. A natural question that arises might be "is there any way to discard some of the possible explanatory variables (that is, shrink their coefficients to zero) *before* implementing a time-consuming LASSO algorithm?" If so, we would need to be able to do this "cheaply" in the computational sense, and we would only want to discard a predictor if we were confident it would be removed via the LASSO procedure anyway.

It turns out that a rule can be built using only inner products of rows of the design matrix $\mathbf{X}$ and the responses, $\mathbf{y}$. For a pre-specified value of λ, we can discard a predictor $\mathbf{x}_j$ if:

$$|\mathbf{x}_j^T\mathbf{y}| < \lambda - ||\mathbf{x}_j||_2||\mathbf{y}||_2\left(\frac{\lambda_{max} - \lambda}{\lambda_{max}}\right)$$

where

$$\lambda_{max} := \max_j |\mathbf{x}_j^T\mathbf{y}|$$

is the smallest value of λ that will shrink all coefficients to zero. For a more in depth look, consult Chuck Vollmer in STAT 007 or read [El Ghaoui et al., 2010]

1 Low Assumption Prediction

Recall, we are taking the low assumption perspective, where we evaluate our estimator based on how it performs relative to the "oracle" estimator, which may or may not match the true $\mathbb{E}[Y|X]$. Define the oracle estimator

$$\beta_t^* = \underset{\{\beta:||\beta||_1\leq t\}}{\arg\min}\ R(\beta)$$

and the *excess risk* as

$$\mathcal{E}(\hat{\beta}_t, \beta_t^*) = R(\hat{\beta}_t) - R(\beta_t^*).$$

If we have a good estimator, then this excess risk should be small. We say a procedure is *persistent* for a set of measures $\mathcal{P}$ if

$$\forall \mathbb{P} \in \mathcal{P}, \qquad \mathcal{E}(\hat{\beta}_t, \beta_t^*) \xrightarrow{p} 0$$

1.1 Persistence Theorem

For one particular family of distributions, $\mathcal{P}$, there is a persistence theorem due to [Greenshtein et al., 2004].

Define the following set of distributions on $\mathbb{R} \times \mathbb{R}^p$: Let $C_{\mathcal{P}} < \infty$ and

$$\mathcal{P} = \{\mathbb{P} : \mathbb{P}Y^2 < C_{\mathcal{P}}, \text{ and } |x_j| < C_{\mathcal{P}} \text{ almost surely, } j = 1, \ldots, p\}$$

This family is one in which the support for each x_j is effectively limited to a subset of $(-C_{\mathcal{P}}, C_{\mathcal{P}})$ (in other words, the marginal tails are chopped off at some finite value), and we're also limiting what happens in the tails of Y marginally. Placing these conditions on the tails makes the proof of this theorem reasonable, but it actually turns out these restrictions are stronger than necessary.

Theorem 1.1. *Over any $\mathbb{P}$ in $\mathcal{P}$, the procedure*

$$\underset{\beta \in \{\beta : ||\beta||_1 \leq t\}}{\arg\min} \quad \hat{\mathbb{P}}\ell_\beta$$

is persistent provided $\log p = o(n)$ and

$$t = t_n = o\left(\left(\frac{n}{\log p}\right)^{1/4}\right)$$

2 Asymptotic Notation

There isn't much to add beyond what's already in Darren's slides here, except answers to the equations he poses on slide 36:

- $o_p(1) + O_p(1) = O_p(1)$
- $o_p(1)O_p(1) = o_p(1)$
- $o_p(1) + o_p(1)O_p(1) = o_p(1)$

3 Inequalities

The proof of the persistence theorem uses several commonly employed inequalities:

- Hölder's:

  For $p, q \in (1, \infty)$ and $\dfrac{1}{p} + \dfrac{1}{q} = 1$,

  We have for all measureable functions f and g

  $||fg||_1 \leq ||f||_p ||g||_q$.

  In particular (and probably more familiarly), for measureable X and Y on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we get:

  $$\mathbb{E}[XY] \leq \mathbb{E}[|X|^p]^{1/p}\, \mathbb{E}[|Y|^q]^{1/q}$$

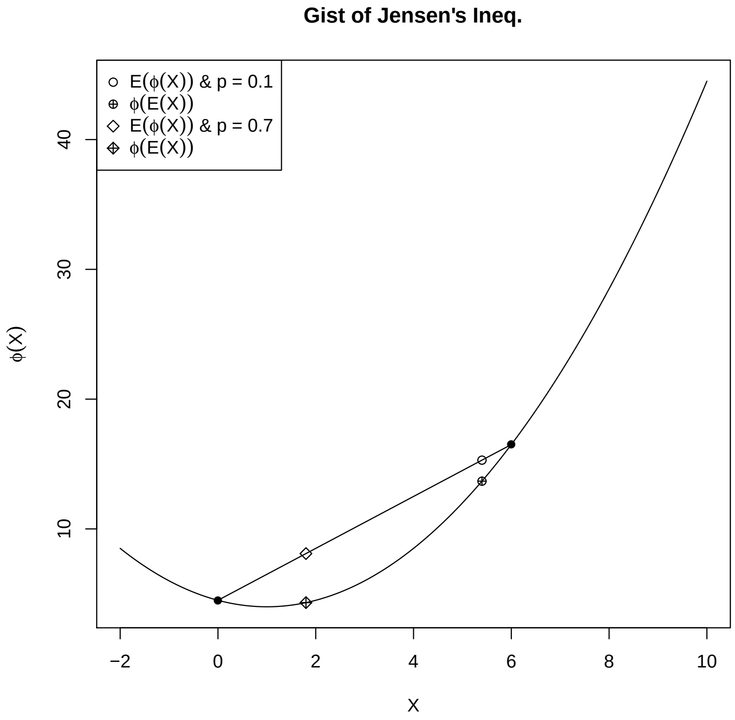

Figure 1: $X \in \{0, 6\}$, $p = 0.1$, and $\phi(X) = 1/2(X-1)^2 + 4$

- Jensen's: For a random variable X and convex function ϕ,

  $$\phi(\mathbb{E}[X]) \leq \mathbb{E}[\phi(X)]$$

  There is a helpful diagram of Jensen's inequality. Consider the simplest case where X is a random variable that takes on one of two values $\{X_1, X_2\}$ with probabilities p for the first value, and $1-p$ for the second. In this case, the value $\phi(\mathbb{E}[X])$ will lie on the curve of $\phi(X)$, and the value $\mathbb{E}[\phi(X)]$ will lie on the line segment which connects $(X_1, \phi(X_1))$ and $(X_2, \phi(X_2))$ **no matter what we choose for** p, and we can clearly see the inequality in action.

  In Figure 1 $X \in \{0, 6\}$, $p = 0.1$ or 0.7, and $\phi(X) = 1/2(X-1)^2 + 4$. No matter what we chose for p, the two points indicating the two sides of the inequality slide together left or right along their respective curves. Thus, the inequality holds no matter what distribution we choose for X. For more complex distributions on X, this diagram becomes more complicated, but at the very least it can help remind us which way the inequality points.

- Markov: If X is any nonnegative integrable random variable and $a > 0$, then

  $$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

  A natural corollary of this is Chebychev's inequality:

  $$\mathbb{P}((X - \mathbb{E}[X])^2 \geq a^2) \leq \frac{\mathrm{Var}[X]}{a^2}.$$

4 Proof of Persistence Theorem

Here are a few short notes explaining some steps of the proof given in the notes:

Line	Justification/Explanation/Clarification/Obfuscation
(32)	add and subtract sample covariance
(33)	since $\hat{\gamma}_t^\top \Sigma \hat{\gamma}_t$ is the minimum value attained over all γ, we can substitute any other value (in this case γ_t^*) to get at least as great a value.
(35)	both $\hat{\gamma}_t$ and $\hat{\gamma}_t^*$ live in the $(t-1)$-ball.
(39)	the max norm is defined $\|\|x\|\|_\infty := \max_i(\|x_i\|)$

Table 1:

Additionally, Nemirovski's Inequality is used, and during it's introduction Darren mentions Banach and Hilbert spaces. The basic difference between these two spaces is that Hilbert spaces come equipped with an inner product, while Banach spaces have only a norm.

5 PAC bounds

At the end of class we mentioned Probably Approximately Correct bounds, where we are trying to covert from

$$\mathbb{P}(\text{error} > \delta) \leq \epsilon$$

to

$$\text{with probability 1 - } \epsilon, \text{ the error is} < \delta.$$

References

Laurent El Ghaoui, Vivian Viallon, and Tarek Rabbani. Safe feature elimination in sparse supervised learning. *CoRR*, 2010.

Eitan Greenshtein, Ya'Acov Ritov, and others. Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 10(6):971–988, 2004. URL http://projecteuclid.org/euclid.bj/1106314846.