Statistical Machine Learning | Lecture 15— Persistence | Oct. 21, 2014

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

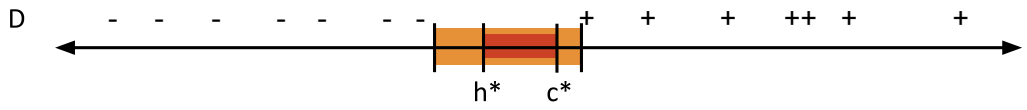
0 A short discussion on Bayesian Lasso - Lei Yang

Lei took a moment to broach the Bayesian perspective on regularization, especially when it comes to an $L1$ "Lasso"–type penalty. It turns out that regularization can be interpreted quite naturally in the Bayesian setting as specifying informative priors on $\boldsymbol{\beta}$. For the case of "Lasso", the priors on $\boldsymbol{\beta}$ take the form of zero-centered Laplace distributions, where the common scale parameter takes the place of the tuning parameter we usually denote by λ. Selecting a reasonable value for this scale parameter can be done through cross validation (either by setting aside data for this purpose, or adopting an empirical Bayes approach), or by choosing a diffuse prior for the scale parameter.

1 PAC learning

The remainder of the lecture is well documented in the slides, with the possible exception of the strong learner example given on pages 6-10, which deserve some accompanying figures. Hopefully the attached explanation clarifies the details of this example, at least up until the conclusion on page 11.

Based on the following example data D, we would have to choose an h* in the orange region.

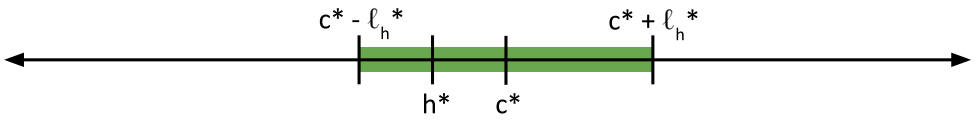


ℓ_h* := length of orange region

Once h* is fixed, the probability of making an error is the probability of drawing a new x from the red region. However, since we don't know where c* is, we bound this region with the green region as follows:

$$P(err|h^*) \leq P([c^* - \ell_h^*, c^* + \ell_h^*])$$

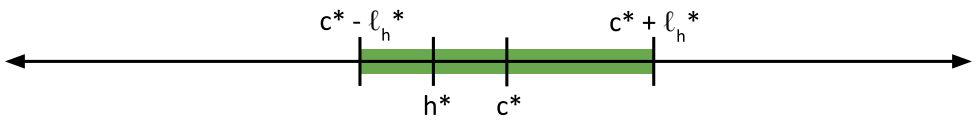
since this covers even the worst case scenario of the red region equalling the orange region.



The probability of making an error greater than ε based on the formulation of a rule h* using training set D of size n will be bounded by

$$P(err > \varepsilon) \leq P^n(P([c^* - \ell_h^*, c^* + \ell_h^*] > \varepsilon | h^*)),$$

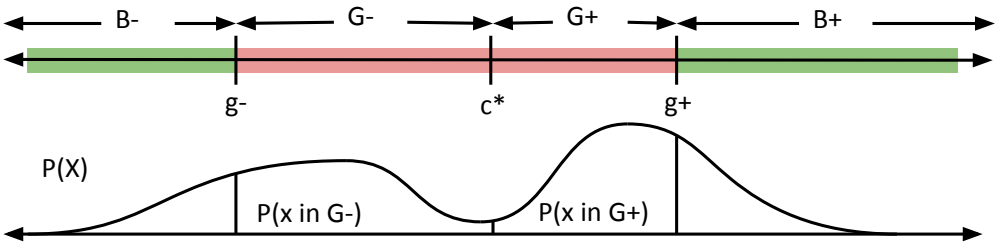
since we can think of first generating h* based on D of size n, and then selecting a new x from the potentially bad region $[c^* - \ell_h^*, c^* + \ell_h^*]$. In order for us to make an error greater than ε, this bad interval must have measure larger than the allowed ε.



Continuing on with some conservative boundaries we obtain the new bound:

$$\begin{aligned} P(err > \varepsilon) &\leq P^n(P([c^* - \ell_h^*, c^* + \ell_h^*] > \varepsilon | h^*)) \\ &\leq P^n(P([c^* - \ell_h^*, c^*] > \varepsilon) \cup P([c^*, c^* + \ell_h^*] > \varepsilon)) \\ &\leq P^n(P([c^* - \ell_h^*, c^*] > \varepsilon)) + P^n(P([c^*, c^* + \ell_h^*] > \varepsilon)) \end{aligned}$$

Now we will further bound these two pieces by considering the following picture, in which we deal with the probability measure on X.



choose the g+ and g- closest to c* such that $P(x \text{ in } G\pm) \geq \varepsilon$

We can bound $P^n(P([c^*, c^* + \ell_h^*] > \varepsilon))$ by making h* come from B+. That is,

$$P^n(P([c^*, c^* + \ell_h^*] > \varepsilon)) \leq P^n(h^* \text{ in } B+)$$

If h* is *not* in B+, then h* < g+ and so $P([c^*, c^* + \ell_h^*])$ may be less than the enforced ε. This is all true for the $[c^* - \ell_h^*, c^*]$ too, of course. Now we bound *this* term, realizing from the first picture that our data determine the max and min value that h* can take on and still be consistent. Namely, all the x_i must come from outside the G+ region or else h* would be less than g+.

$$\begin{aligned} P^n(h^* \text{ in } B+) &\leq P(x_1 \notin G+) \ldots P(x_n \notin G+) \\ &\leq (1 - \varepsilon)^n \\ &\leq e^{-\varepsilon n} \end{aligned}$$

Now we have

$$P(\text{err} > \varepsilon) \leq 2e^{-\varepsilon n}$$