Model Selection for Small Sample Regression

Olivier Chapelle*, Vladimir Vapnik*,†, Yoshua Bengio*,††
{*chapelle,vlad,yoshua*}@*research.att.com*
* AT&T Research Labs
† Royal Holloway, University of London, UK
†† Université de Montréal, Canada

1 Introduction

Model selection is an important ingredient of many machine learning algorithms, in particular when the sample size in small, in order to strike the right trade-off between overfitting and underfitting. Previous classical results for linear regression are based on an asymptotical analysis. We present a new penalization method for performing model selection for regression that is appropriate even for small samples. Our penalization is based on accurate estimate of the ratio of the expected training error and the expected generalization error, in terms of the expected eigenvalues of the input covariance matrix.

2 Risk of the mean square estimor

Given a collection of data $(x_1, y_1), ..., (x_n, y_n)$, where $y_i = f(x_i, \alpha_0) + \xi_i$ and x_i, ξ_i are independently generated by unknown distributions $P(x)$ and $P(\xi)$, one wants to find the function $\hat{f}$ that provides the smallest value of the expected loss $E_{x,\xi}(y - \hat{f}(x)^2)$

We consider a linear model of dimension d,

$$\mathcal{F}_d = \left\{ x \to \sum_{i=1}^{d} \alpha_i \varphi_i(x) \right\} \tag{1}$$

with $\alpha_i \in \mathbb{R}$ and the family $\{\varphi_i(x)\}_{i \in \mathbb{N}}$ is orthonormal with respect to the probability measure $\mu(x)$, which means $E\ \varphi_p(x)\varphi_q(x) = \delta_{pq}$. [1]

The choice of a good dimension d is called the model selection problem and is at the heart of the bias / variance dilemna. Akaike's (FPE) (Akaike, 1970) suggests to choose the model d which minimizes

$$\sum_{i=1}^{n} \left(y_i - \hat{f}_d(x_i)\right)^2 \left(1 + \frac{d}{n}\right) \left(1 - \frac{d}{n}\right)^{-1}$$

Other model selection criterion have the same asymptotical multiplicative correcting term in $(1 + 2d/n)$.

[1] Note that the choice of such a family requires knowledge about $P(x)$. We will see later how to avoid this assumption

Most of the correcting terms proposed in the litterature have been proven to be asymptotically efficient but are rarely used in practice because they behave poorly for small sample size and cross-validation method is usually prefered.

We propose an approach which corrects Akaike's term for small sample size. More precisely our main result is the following:

$$\boxed{ER(\hat{f}_d) \approx ER_{emp}(\hat{f}_d)\left(1 - \frac{d}{n}\right)^{-1}\left(1 + \frac{E\sum_{i=1}^{d}(1/\lambda_i)}{n}\right)} \tag{2}$$

As before $\hat{f}_d$ is the function minimzing the empirical risk on the model of dimension d, R and R_{emp} are the associated risk and empirical risk respectively and λ_i are the eigenvalues of the empirical covariance matrix,

$$(C_{emp})_{pq} = \frac{1}{n}\sum_{i=1}^{n}\varphi_p(x_i)\varphi_q(x_i).$$

As the goal in model selection is to choose the model with the smallest expected risk, the previous analysis suggests to take the model d which minimizes the right hand-side of equation (2)

Note that in the asymptotic case, since the covariance matrix is almost the identity matrix (from the orthonormality assumption), $E(1/\lambda_i) \approx 1$ and we obtain Akaike's term. However, in the non-asymptotic case the covariance matrix is not well-conditioned and it can happen that $E(1/\lambda_i) >> 1$

We made this assumption that the basis should be orthonormal with respect to the input distribution $P(x)$ which requires some prior knowledge about it. In the case this distribution is unknown but along with training data, "unlabeled" data are available (x without y), it is still possible to estimate the correcting term. One can compute two covariance matrices: one from unlabeled data $\tilde{C}$ and an other from the training data C_{emp} and get

$$\sum_{i=1}^{d}(1/\lambda_i) \approx \text{trace } (C_{emp}^{-1}\tilde{C}),$$

which is used to evaluate equation (2).

Experimental results show that this technique outperforms other penalty based model selection methods ans is comparable to the state of the art (Schuurmans, 1997).

References

Akaike, H. (1970). Statistical predictor identification. *Ann. Inst. Stat. Math.*, 22:202–217.

Schuurmans, D. (1997). A new metric-based approach to model selection. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence*.