CS281B/Stat241B: Advanced Topics in Learning & Decision Making

Cross Validation, Regularization, and Information Criteria

Lecturer: Michael I. Jordan *Scribes: Charless Fowlkes & Bhaskara Marthi*

1 Cross Validation

Cross validation is one of the simplest and often most practical approaches to model selection. The setup is that we are given training data $\mathcal{D}$ and test data $\mathcal{D}_{test}$, and we would like to fit this data with a model $p(x|\theta, h)$ from the family $\mathcal{F}$ which is indexed by h and θ. h is understood to specify 'structural' properties of the model, while θ specifies parameter values. In practice the boundary between these two is somewhat fuzzy.

The most common form of cross-validation is known as K-fold cross-validation. It works as follows :

- Set aside αN samples of $\mathcal{D}$ (where $N = |\mathcal{D}|$). This is known as the held-out data and will be used to evaluate different values of h.
- For each candidate structure h, fit $p(x|\theta, h)$ to the remaining $(1-\alpha)N$ samples in $\mathcal{D}$ (i.e hold h fixed and find the best θ).
- Now that we have found a parameter value θ_h corresponding to each h, evaluate each model $p(x|\theta_h, h)$ on the held-out data using some pre-specified risk function.
- Repeat the above K times, choosing a different held-out data set each time, and the results are averaged for each h. This gives an estimate of the risk curve.
- Let h^* denote the value of h with the best overall performance. Use all of $\mathcal{D}$ to find the parameter values for $p(x|\theta, h^*)$.

One important point is that the test data $\mathcal{D}_{test}$ is never used in the above process, because doing so would result in overly optimistic accuracy rates during the testing phase.

Commonly used values for K and α are $K = 10$ and $\alpha = 0.1$.

There are some problems with this approach. First, it's not viable when data sets are small relative to the number of models that are being evaluated. Second, K needs to be large for the variance to be small enough, and this makes it time-consuming.

2 An Application to Histograms

Cross-validation can be used to find the optimal number of bins for a histogram model. For a given value of h, the bin width, the integrated squared error is :

$$ISE_h = \int (\hat{f}(x) - f(x))^2 dx$$

$$
\begin{aligned}
&= \int \hat{f}(x)^2 dx - 2\int \hat{f}(x) f(x) dx + \int f(x)^2 dx \\
&= \int \hat{f}(x)^2 dx - 2E(\hat{f}(x)) + C \\
&\approx \int \hat{f}(x)^2 dx - \frac{2}{N}\sum_{1}^{N} \hat{f}_{-n}(x_n) + C.
\end{aligned}
$$

where C is constant with respect to h and can be ignored. In the second term, $\hat{f}_{-n}$ is the histogram value at x_n for the histogram generated with the data $\{\mathcal{D} - \{x_n\}\}$; i.e., this is a leave-one-out estimate. As we show in Homework 1, there is a closed-form expression for this sum; i.e., these estimates can be computed from a single pass through the data.

3 Regularization

In the early 60's, James and Stein showed that the maximum-likelihood estimate of the mean of a diagonal multivariate gaussian is inadmissible in $\Re^N$ for $N > 3$, i.e there exists no parameter value for which the ML estimate is the best one. The proof is constructive—it yields the James-Stein estimator which is strictly better (lower risk) everywhere in the parameter space.

A class of alternatives to ML inspired by the James-Stein result is the regularization methods (also known as penalized likelihood, shrinkage, and smoothing methods). In all of these are methods one optimizes a loss function that takes the form of a penalty term added to the likelihood. For example,

$$\hat{\theta}_{shrinkage} = \text{argmax}_\theta \left[l(\theta|\mathcal{D}) - \lambda \|\theta\| \right]$$

where $\lambda > 0$ and $\|\theta\|$ might be the L_1 or L_2 norm. The choice of norm has an effect—using the L_2 norm pulls θ directly towards the origin, while using the L_1 norm pulls towards the coordinate axes, i.e it tries to set some of the coordinates to 0. This second approach can be useful in a feature-selection setting.

Also, $\lambda\|\theta\|$ can be replaced $\lambda\|\theta - a\|$ for some fixed a. However, by convention, a is set to 0.

The reason for using such estimators is that, while the estimate is no longer unbiased, it has a lower variance (since it shrinks things towards the origin), and may therefore reduce the overall risk.

An example in text processing: modeling the probability of a document by the probability of each word contained in it. Maximum likelihood would assign all of the probability to words seen during training. So any document containing an unseen word has a probability of 0. Adding a shrinkage term reserves some of the probability for unseen words.

Another example is linear regression. Recall that using squared error as the cost function results in the least squares estimate, which is unbiased. However, we can also use

$$l = -\frac{1}{2}\sum_n (y_n - \theta^T x_n)^2 - \lambda \sum_i \theta_i^2$$

hoping that the increased bias is counterbalanced by the reduction in variance. Regularizing l in this way yields a least square problem solved by a "corrected" version of the normal equations.

One question about these methods is how to choose λ. One generally uses cross-validation.

4 Information Criteria

A traditional question in linear regression is how many features to use. Again, ML is inappropriate as it would favor using an arbitrarily large number of features. One approach is is to maximize $l(\theta|\mathcal{D}) - k$ where k is the number of parameters. This particular penalty term emerges from an approach to model selection known as the AIC (An Information Criterion), or more generally the TIC (Takeuchi Information Criterion).

The AIC was developed by Akaike. A significant point of departure from classical approaches to estimation is that it does not assume that the true density $f(x)$ belongs to the family of candidate models $\mathcal{F} = \{g(x|\theta)\}$.

For a given family, let:

$$\theta_0 = \text{argmin}_\theta D\{f(x)\|g(x|\theta)\}$$

θ_0 corresponds to the member of $\mathcal{F}$ that is closest to f, using KL divergence as a measure of "closeness."

We would like to estimate θ_0. Suppose we have sampled y from $f(x)$. For the given model family and data, we can compute a maximum likelihood estimate which we denote as a function $\hat{\theta}_{ML}(y)$ of these data samples.

Now the divergence

$$D\{f(x)\|g(x|\hat{\theta}_{ML}(y))\}$$

is a function of y and hence it is a random variable. We would like to minimize the expected value of this quantity which, by definition of KL divergence, is the same as maximizing:

$$E_y \int f(x) \log g(x|\hat{\theta}_{ML}(y))dx = E_y E_x \log g(x|\hat{\theta}_{ML}(y)).$$

Note that in maximizing this quantity we are taking into account the fact that a given model has a certain variability in the maximum likelihood estimator $\hat{\theta}_{ML}$. We try to minimize a KL divergence that takes this variability into account. This is different from simply trying to "estimate θ_0."

Having maximized the expression in Equation **??** for each model g in $\mathcal{F}$, we choose the model which maximizes this value.

However there is still the question of how to calculate this value. In particular, neither expectation can be directly evaluated. So we will have to approximate it somehow.

5 References:

- Akaike, H. Information theory and an extension of the maximum likelihood principle. Reprinted in Samuel Kotz and Norman L. Johnson (eds.), *Breakthroughs in Statistics*, Volume I. New York: Springer-Verlag, 1992.
- Efron, B. and Tibshirani, R. *An Introduction to the Bootstrap*. New York: Chapman and Hall, 1993.
- Shibata, R. An optimal selection of regression variables. *Biometrika*, 68(1):45-54, 1981.