Statistical Machine Learning: Clustering and Dimension Reduction

Daniel J. McDonald

Indiana University, Bloomington

mypage.iu.edu/∼dajmcdon

February 24-26, 2015

Lessons from this course

Predictive risk

Suppose I come up with some method of description / estimation / prediction. . .

Is it any good?

How can I evaluate my method?

Are there better methods?

Evaluate methods using predictive risk

What if I wanted to make inferences?

You think you don’t care about predictions.

You want to describe the world. This is admirable.

If your model for the world makes predictions that are inaccurate, then your model is not very good. Popper’s falsification principle for inductive reasoning

Good predictive performance is a necessary condition for me to believe your description.

Prediction risk

We evaluate prediction quality with the prediction risk.

Choose a loss function that correctly measures the cost of poor predictions in the problem at hand.

Prediction risk

$$R_n(\widehat{f}) = \mathbb{E}[\ell(Y, \widehat{f}(X))]$$

where the expectation is taken over the new data point (Y, X) and $\mathcal{D}_n$ (everything that is random).

Bias-variance decomposition

In many cases,

$$\text{prediction risk} \quad = \quad \text{bias}^2 \quad + \quad \text{variance} \quad + \quad \text{irreducible error}$$

Prediction risk is proportional to estimation risk. However, defining estimation risk requires stronger assumptions.

In order to make good predictions, we want our prediction risk to be small. This means that we want to 'balance' the bias and variance.

Bias-variance tradeoff

- bias: how well does $\widehat{f}$ approximate the truth g
- more complicated $\mathcal{F}$, lower bias. Flexibility $\Rightarrow$ Parsimony
- more flexibility $\Rightarrow$ larger variance
- complicated models are hard to estimate precisely for fixed n
- irreducible error

Tuning parameters

Tuning parameters in regularized algorithms trade bias and variance.

More regularization $\Rightarrow$ More bias, less variance

Less regularization $\Rightarrow$ Less bias, more variance

Somewhere in the middle is the predictor with the smallest risk.

Estimating risk

Training error is a bad estimator of risk.

It is optimistic.

Cross validation works better.

Use your risk estimator to choose tuning parameters to balance bias and variance.

Wrapping it all up

Bias is good

1. All models are mis-specified
2. Mis-specified models may still predict well
3. Good predictive risk comes from balancing bias and variance
4. Very often, we can trade some bias for (much) lower variance
5. Bias is controlled by setting tuning parameters
6. Choosing tuning parameters carefully gives good risk properties
7. To know how to choose the tuning parameters, we need an estimate of the risk
8. Training error (which a risk estimator) is a bad choice (optimistic)

Thanks for listening.