Statistics 452: Statistical Learning and Prediction

Review of Key Ideas

Brad McNeney

2017-11-30

Focus on Supervised Learning

- Our focus (chapters 3-9) was on supervised learning, where there is a response to validate our models.
- Won't discuss unsupervised learning in this review.

Models

- The general model for a response Y is

$$Y = f(X) + \epsilon$$

  where

  - f is a fixed but unknown function that is the **systematic** component of the model
    - We usually take $f(X)$ to be the mean of Y given X.
  - ϵ is an error component, assumed to be independent of X and to have mean zero.
    - Even if Y is, say, binary, the errors have mean zero.
- We studied different approaches for
  - estimating f and
  - quantifying the accuracy of the estimate

Goals of Estimation

1. prediction
2. inference

Prediction

- Since the errors average to zero, $f(X)$ is a reasonable prediction of a new Y.
- Based on and estimate $\hat{f}$ of f the estimate, or prediction of Y is

$$\hat{Y} = \hat{f}(X)$$

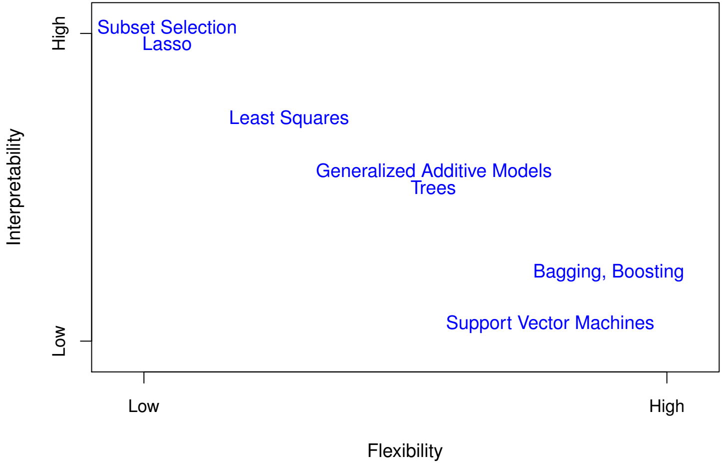
- One school of thought treats $\hat{f}$ as a "black box".
  - We do not really care about the details of $\hat{f}$, only that its predictions $\hat{Y}$ are accurate.
  - Among statisticians, the chief proponent of this view was Leo Brieman.

Inference

- Or, should our goal be to "open the box" and see what's inside?
  - See first 4:30 of TED talk by Barbara Englehardt: `https://www.youtube.com/watch?v=uC3SfnbCXmw`
  - Reference: Brieman (2001). Statistical Modeling: The Two Cultures. Copy available on Canvas
- Classically, inference means inference of parameters in simple parametric models for f.
  - Could also include nonparametric methods such as smoothing splines, parametrized by a df, and for which df=1 is a linear model.

Flexibility *versus* Interpretability

- Most methods can be used for **both** prediction and inference; i.e., can't be classified strictly as closed or open box.
  - Can rate methods in terms of flexibility, which comes at the cost of interpretability. Schematically (text, Fig. 2.7):



Model Accuracy

Loss Functions

- Reference: Elements of Statistical Learning, Chapter 7.
- We measure the errors between Y and fit $\hat{f}(X)$ by a loss function $L(Y, \hat{f}(X))$.
- For quantitative Y we have used squared error loss

$$L(Y, \hat{f}(X)) = (Y - \hat{f}(X))^2$$

- For categorical response, G, we have not been explicit about loss functions, but have mentioned zero-one loss (misclassification error), logistic loss (logistic regression) and others.

Training Error

- The training error is the average loss over the training set.
- For example, using squared error loss

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 = \overline{\mathrm{err}} \tag{1}$$

The Test Error

- For fixed x_0 and y_0 **not used to train** $\hat{f}$, the expected test error is $E(L(Y, \hat{f}(X)|y_0, x_0)$.
- For example, with square error loss we have the test MSE $E(y_0 - \hat{f}(x_0))^2$, obtained by averaging over training data sets (repeated estimations of f).
- Test MSE can be decomposed as

$$E(y_0 - \hat{f}(x_0))^2 = Var(\hat{f}(x_0)) + [Bias(\hat{f}(x_0))]^2 + Var(\epsilon)$$

  where

  - $Var(\hat{f}(x_0))$ is the variance (spread) of the predictions,
  - $Bias(\hat{f}(x_0))$ is the bias (systematic departure from truth) of the predictions, and
  - $Var(\epsilon)$ is the irreducible error term that is beyond our control

- The overall average test MSE, $\text{Err}_{\mathcal{T}}$is calculated by averaging over a test set $\mathcal{T}$ of (x_0, y_0) pairs.

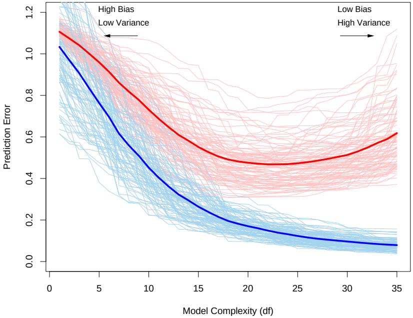


FIGURE 7.1. *Behavior of test sample and training sample error as the model complexity is varied. The light blue curves show the training error* $\overline{\text{err}}$*, while the light red curves show the conditional test error* $\text{Err}_{\mathcal{T}}$ *for* 100 *training sets of size* 50 *each, as the model complexity is increased. The solid curves show the expected test error* Err *and the expected training error* $\text{E}[\overline{\text{err}}]$.

Bias-Variance Tradeoff

- Generally, the more flexible the method for estimating f the higher the variance and the lower the bias.
  - Initially as we increase flexibility, the variance increase is offset by a decrease in bias, and the test MSE decreases.
  - At some point though the variance increase exceeds the decrease in bias and the expected test MSE increases.
- Flexibility can be increased by adding more predictors, powers of predictors, basis functions, etc.
- Flexibility can be reduced by restricting model terms or shrinking coefficients.

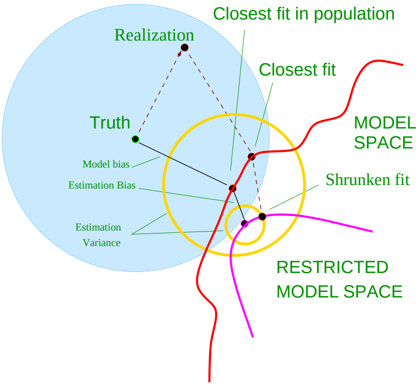


FIGURE 7.2. *Schematic of the behavior of bias and variance. The model space is the set of all possible predictions from the model, with the "closest fit" labeled with a black dot. The model bias from the truth is shown, along with the variance, indicated by the large yellow circle centered at the black dot labeled "closest fit in population." A shrunken or regularized fit is also shown, having additional estimation bias, but smaller prediction error due to its decreased variance.*

Estimating Accuracy

Estimated Test MSE

- If the training observations $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ are used to produce $\hat{f}$, and we had a large number of test observations (x_0, y_0), the test MSE

$$Ave(y_0 - \hat{f}(x_0))$$

reflects how well $\hat{f}$ predicts new observations.
- We would like to develop methods that minimize the test MSE.
- Validation and cross-validation (CV) are tools to estimate the test MSE.

Validation and Cross-Validation

- Validation: Split the data into two parts, a training set and a validation, or hold-out set.
  - Use the training set for fitting and the validation set for estimating the test error.
- Cross-Validation (CV): Split the data into multiple "folds" of approximately equal size.
  - Common numbers of folds are $k = n$, 10 and 5.
  - Train on all but one hold-out fold, and test on the hold-out to get MSE_i; $i = 1, \ldots, k$. Repeat for each fold and average the estimated test MSEs:

$$CV_{(k)} = \frac{1}{k} \sum_{i=1}^{k} \mathrm{MSE}_i.$$

Estimation

Common Themes in Estimation

- Parametric *versus* non-parametric forms for f
- Additive models and basis functions.
- Model selection or shrinkage to control flexibility (complexity).

Parametric Methods

- Specify a form for f that depends on a finite number of parameters, and estimate these parameters by minimizing a criterion function for training data.
  - Examples include least squares or penalized least squares regression.
- The true f may not be well-approximated by the functional form we choose for our parametric model.
- We can choose a very flexible parametric family, but if too flexible we may **overfit**; i.e., the fitted model may follow the error terms.

Non-parametric Methods

- An model-free specification of the functional form of f, fit to the training data.
  - Examples include smoothing splines and KNN.
- Avoid over-fitting by limiting the roughness, or wigglyness of the fitted curve.
  - E.G., df of the smoothing spline, neighborhood size for KNN.
- Non-parametric methods require more data that a parametric method to train the model to obtain accurate estimates.

Additive Models and Basis Functions

- A general additive models is of the form

$$f(x;\alpha,\gamma) = \sum_{m=1}^{M} \alpha_m b(x;\gamma_m)$$

  for coefficients α and basis function parameters γ
- We studied linear and logistic regression with basis functions such as power, and piecewise-cubic splines.
- Generalized additive models can use local regression or smoothing spline basis functions.
- Boosting uses decision trees as basis functions.

Model Selection

- Select the number of model terms that minimizes the test error, estimated by CV.
- We typically don't consider all possible models, but rather choose a search strategy, such as forward stagewise selection.

Minimize Loss Plus Complexity

- The lasso and ridge regression minimize squared error or logistic loss, plus a tuning parameter times an ℓ_1 or ℓ_2 complexity penalty.
- SVM: hinge loss plus tuning parameter times ℓ_2 penalty.
- Select the tuning parameter by CV

Curse of Dimensionality

- We may think that more predictors is a good thing, but too many predictors that are unrelated to the response lead to poor performance.
- Referenced most often for KNN
- Also saw that shrinkage methods performed poorly when there are many useless predictors.