Lecture 10 – Model Assessment and Selection

Martin Spindler

2016-05-04

Bias, Variance, and Model Complexity

- Target variable Y, inputs X, $\hat{f}(X)$ prediction model estimated from training set $\mathcal{T}$
- Typical choices of loss functions: $L(Y, \hat{f}(X)) = (Y - \hat{f}(X))^2$ (squared error) or $L(Y, \hat{f}(X)) = |Y - \hat{f}(X)|$ (absolute error)
- The test error / generalization error, is the prediction error over an independent test sample

$$Err_{\mathcal{T}} = E[L(Y, \hat{f}(X))|\mathcal{T}]$$

where both X and Y are drawn randomly from their joint distribution (population).

Bias, Variance, and Model Complexity

- Expected prediction error (or expected test error)

$$Err = E[L(Y, \hat{f}(X))] = E[Err_{\mathcal{T}}]$$

Bias, Variance, and Model Complexity

- Goal: estimation of $Err_{\mathcal{T}}$
- Training error is the average loss over the training sample:

$$\bar{err} = 1/n \sum_{i=1}^{n} L(y_i, \hat{f}(x_i)).$$

- Similar for categorical variables (but different loss function).

Bias, Variance, and Model Complexity

- Usually model depends on a tuning parameter α: $\hat{f}_\alpha(x)$
- Two different goals:
  - Model selection: estimating the performance of different models in order to choose the best one.
  - Model assessment: having chosen a final model, estimating its prediction error (generalization error) on new data.

The Bias-Variance Decomposition

- $Y = f(X) + \varepsilon$ with $E[\varepsilon] = 0$ and $Var(\varepsilon) = \sigma_\varepsilon^2$
- The expected prediction error of a regression fit $\hat{f}(X)$ at a point $X = x_0$ under squared-error loss is given by

$$Err(x) = \sigma_\varepsilon^2 + Bias^2(\hat{f}(x_0)) + Var(\hat{f}(x_0))$$

- This can be interpreted as "Irreducible Error + Bias^2 + Variance".

The Bias-Variance Decomposition | Example OLS

- For linear model fit $\hat{f}_p(x) = x^T\hat{\beta}$ with p components by ols we have

  $$E(x_0) = E[(Y - \hat{f}_p(x))^2 | X = x_0] = \sigma_\varepsilon^2 + [f(x_0) - E\hat{f}_p(x_0)]^2 + \|h(x_0)\|^2\sigma_\varepsilon^2$$

  $$h(x_0) = X(X^TX)^{-1}x_0$$

The Bias-Variance Decomposition | Example OLS

- Average over all sample values x_i gives:

$$1/n \sum_{i=1}^{n} Err(x_i) = \sigma_\varepsilon^2 + 1/n \sum_{i=1}^{n} [f(x_i) - E\hat{f}(x_i)]^2 + \frac{p}{n}\sigma_\varepsilon^2,$$

  the in-sample error.

Optimism of the Training Error Rate

- With $\mathcal{T} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ given, the generalization error of a model $\hat{f}$ is

$$Err_{\mathcal{T}} = E_{X^0,Y^0}[L(Y^0, \hat{f}(X^0))|\mathcal{T}]$$

  (fixed training set $\mathcal{T}$, new observation /data point (X^0, Y^0) drawn from F, the distribution of the data)
- Averaging over training sets yields the expected error:

$$Err = E_{\mathcal{T}} E_{X^0,Y^0}[L(Y^0, \hat{f}(X^0))|\mathcal{T}]$$

  (easier to analyze)
- In general: $\bar{err} = 1/n \sum_{i=1}^{n} L(y_i, \hat{f}(x_i)) \leq Err_{\mathcal{T}}$

Optimism of the Training Error Rate

- Part of the discrepancy come from the location where the evaluation points occur. $Err_{\mathcal{T}}$ as extra-sample error.
- In-sample error (for analysis of $\bar{err}$)

$$Err_{in} = 1/n \sum_{i=1}^{n} E_{Y^0}[L(Y_i^0, \hat{f}(x_i))|\mathcal{T}]$$

  (observation of n new response values at each of the training points x_i)
- Optimism: difference between Err_{in} and training error $\bar{err}$:

$$op \equiv Err_{in} - \bar{err}.$$

- Average optimism is the expectation of the optimism over training sets:

$$\omega \equiv E_y(op).$$

Optimism of the Training Error Rate

- Usually only ω and not *op* can be estimated (analogous to *Err* and $Err_{\mathcal{T}}$)
- It can be shown: $\omega = 2/n \sum_{i=1}^{n} Cov(\hat{y}_i, y_i)$.
- Interpretation
- In sum: $E_y(Err_{in}) = E_y(\bar{err}) + 2/n \sum_{i=1}^{n} Cov(\hat{y}_i, y_i)$
- Example: linear fit with p variables for model $Y = f(X) + \varepsilon$: $\sum_{i=1}^{n} Cov(\hat{y}_i, y_i) = p\sigma_{\varepsilon}^2$

Estimates of In-Sample Prediction Error

- General form of the in-sample estimates: $\hat{Err}_{in} = \bar{err} + \hat{\omega}$.
- C_p statistic: $C_p = \bar{err} + 2\frac{d}{n}\hat{\sigma}_\varepsilon^2$
- Akaike Information Criterion: $AIC = -\frac{2}{n} loglik + 2\frac{d}{n}$
- Bayesian Information Criterion: $BIC = -2 loglik + (\log n)d$

Cross-Validation

- Estimation of the prediction error directly.
- CV estimates the expected extra-sample error $Err = E[L(Y, \hat{f}(X))]$
- Formal description:
  - Denote κ a partitioning function: $\kappa : \{1, \dots, n\} \to \{1, \dots, K\}$
  - Denote by $\hat{f}^{-k}(x)$ the fitted function, computed with the kth part of the data removed.

Cross-Validation

- The cross-validated estimator of the prediction error is

$$CV(\hat{f}) = 1/n \sum_{i=1}^{n} L(y_i, \hat{f}^{-k}(x_i)).$$

- Typical choices : $K = 5, 10$, $K = n$ is called *leave-one-out* cross-validation

Cross-Validation | Tuning Parameter

Given a set of models $f(x, \alpha)$ indexed by a tuning parameter α, denote by $\hat{f}^{-k}(x, \alpha)$ the model fit with the kth part of the data removed and tuning parameter α. Then for this set of model we define

$$CV(\hat{f}, \alpha) = 1/n \sum_{i=1}^{n} L(y_i, \hat{f}^{-k}(x_i, \alpha)).$$

The function $CV(\hat{f}, \alpha)$ provides an estimate of the test error curve, and we find the tuning parameter $\hat{\alpha}$ that minimizes it. Our final is $\hat{f}(x, \hat{\alpha})$, which we then fit to all the data.