Cross-Validation

Data 100: Principles and Techniques of Data Science

Sandrine Dudoit

Department of Statistics and Division of Biostatistics, UC Berkeley

Spring 2019

Outline

Version: 04/04/2019, 21:27

Risk-Based Inference

- This section of the course on statistical inference is concerned with the broad question of using data to infer/learn relationships among variables.
- E.g. How can we predict rent for apartments in Berkeley?
- E.g. Which features of a car are related to its fuel consumption?
- In regression, the function describing the relationship between an outcome $Y \in \mathbb{R}$ and covariates $X \in \mathbb{R}^J$ is the conditional expected value of the outcome given the covariates, $\theta(X) = \mathrm{E}[Y|X]$.
- As we discussed in earlier lectures, different types of models/estimators can lead to very different fits.

Risk-Based Inference

- In particular, there is a bias-variance trade-off, in the sense that more complex estimators tend to have less bias but more variance than simpler estimators.
- For instance, high-degree polynomial regression functions could overfit the learning data.
- Instead of seeking estimators that simultaneously minimize both bias and variance, one seeks to minimize risk or maximize accuracy, i.e., the average "distance" between an estimator and the parameter of interest.

Regression Example

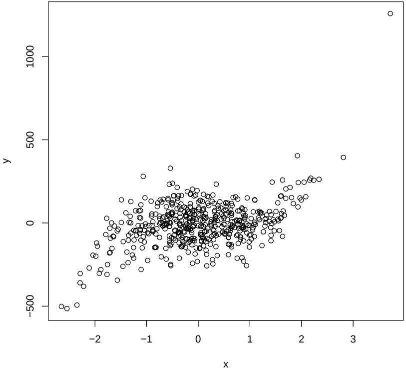

Figure 1: *Regression.* Scatterplot of 500 covariate-outcome pairs from an unknown data generating distribution. What is the regression function?

Regression Example: Model Complexity

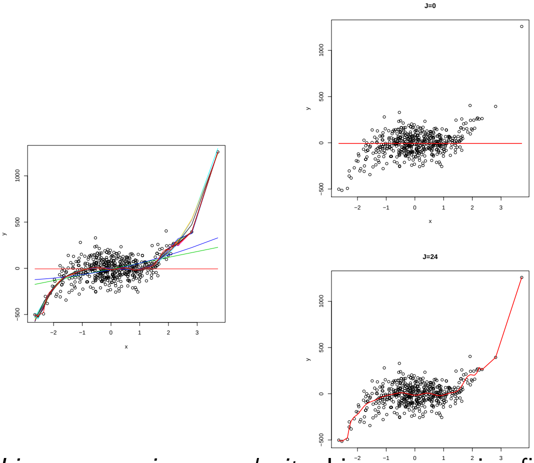

Figure 2: *Linear regression complexity.* Linear regression fits for polynomials of degree 0 to 24.

Regression Example: Model Complexity

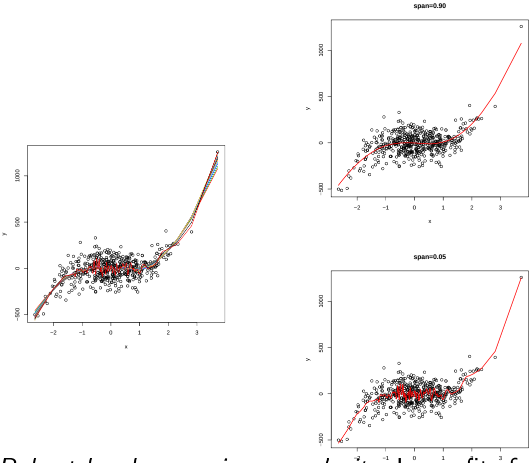

Figure 3: *Robust local regression complexity.* Loess fits for spans ranging from 0.05 to 0.90.

Bias-Variance Trade-Off

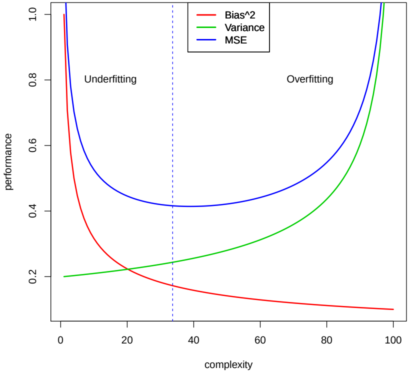

Figure 4: *Bias-variance trade-off.* Schematic representation of bias-variance trade-off as a function of model complexity.

Risk-Based Inference

- Optimal statistical inference. A very broad class of statistical inference methods can be framed in terms of risk optimization.
- Least squares estimation (LSE) involves minimizing risk for the squared error loss function.
- Maximum likelihood estimation (MLE) involves minimizing risk for the negative log loss function.
- In risk-based inference, loss functions and their expected values, i.e., risk functions, are used to
  - identify/select an appropriate model, i.e., a set of distributions or parameters for the population and data generating mechanism;
  - fit the model to the data, i.e., derive an "optimal" estimator of the parameter of interest given the model and data;

Risk-Based Inference

- assess the performance of the model/estimator, cf. accuracy of estimator/prediction.

- In practice, however, one cannot compute the true population risk, i.e., the expected value of the loss function with respect to the population distribution P, as P is unknown.
- Instead, one has to use the available data or learning data to estimate risk.

Learning Set Risk

- Suppose one has a learning set $\mathcal{L}_n = \{(X_i, Y_i) : i = 1, \ldots, n\}$ that is a random sample of n covariate/outcome pairs from the population of interest.
- A naive risk estimator is the learning set risk or resubstitution risk, i.e., the expected value of the loss function with respect to the known data empirical distribution P_n for the learning set in place of the unknown population distribution P.
- The learning set risk is simply the average of the loss function evaluated at each observation in the learning set

$$\frac{1}{n}\sum_{i=1}^{n} L((X_i, Y_i), \theta). \tag{1}$$

Learning Set Risk

- For the squared error loss function used in regression, the learning set mean squared error (MSE) is

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \theta(X_i))^2. \tag{2}$$

- Unfortunately, selecting models/estimators by minimizing learning set risk over large models/parameter spaces leads to overfitting of the learning data, i.e., to estimators that best fit the learning set, but not necessarily an independent test set from the same population.
- Minimizing learning set risk can still lead to accurate estimators, provided risk is minimized over models/parameters spaces that are not too large/complex.

Test Set Risk

- In some cases, one may have access to a test set, i.e., an independent random sample from the same population as the learning set: $\mathcal{T}_n = \{(X_i^*, Y_i^*) : i = 1, \ldots, n^*\}$
- A sensible estimator of the risk for an estimator $\hat{\theta}$ based on the learning set is the test set risk, i.e., the average of the loss function for each of the observations in the test set

$$\frac{1}{n^*} \sum_{i=1}^{n^*} L((X_i^*, Y_i^*), \hat{\theta}). \tag{3}$$

Learning and Test Set Risk

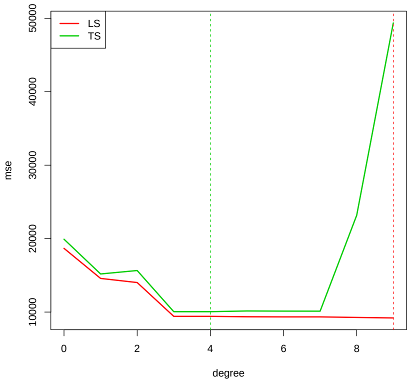

Figure 5: *MSE: Linear regression.* Learning and test set MSE for linear regression fits for polynomials of degree 0 to 9 ($n = 500$, $n^* = 10,000$). Dashed lines indicate risk minimizer.

Learning and Test Set Risk

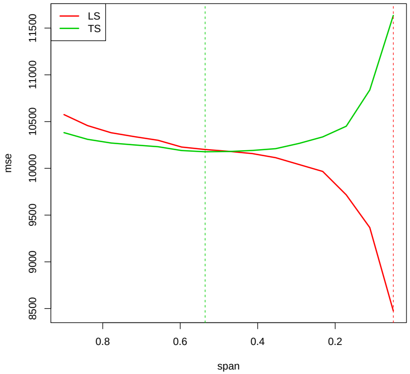

Figure 6: *MSE: Robust local regression.* Learning and test set MSE for loess fits for spans ranging from 0.05 to 0.90 ($n = 500$, $n^* = 10,000$). Dashed lines indicate risk minimizer.

Cross-Validation

- In many cases, however, one does not have access to a test set.
- Instead, we can cleverly divide the learning set into data for training estimators and data for validating their performance, i.e., computing risk.
- This is the main idea behind cross-validation (CV):
  - Partition the available learning set into two sets: A training set and a validation set.
  - Observations in the training set are used to compute, or train, estimators.
  - Observations in the validation set are used to assess the risk of, or validate, the estimators.
- One of the most common forms of cross-validation is K-fold cross-validation.

Cross-Validation

- Randomly partition the learning set into K mutually exclusive and exhaustive sets of approximately equal size.
- Use each of the K sets in turn as a validation set to assess risk for estimators computed using the remaining $(K-1)$ sets as a training set.
- The cross-validated risk estimator is the average of the K validation set risks.
- Smaller values of the number of folds K tend to lead to lower variance (larger validation set), but higher bias (smaller training set) in risk estimation.
- Common choices for the tuning parameter K are between 5 and 10.

Cross-Validation

- Another common type of cross-validation is Monte-Carlo cross-validation, where the learning set is repeatedly randomly partitioned into a training set comprising $(1-\kappa)100\%$ of the learning set and a validation set comprising the remaining observations. Common values for κ are between 0.05 and 0.20.
- When using cross-validation for model selection, e.g., selecting the degree of a polynomial or features to include in a regression model, we select the model with lowest cross-validated risk.
- In order to assess the performance of the final selected model, we should use, if available, an independent test set.

Cross-Validation

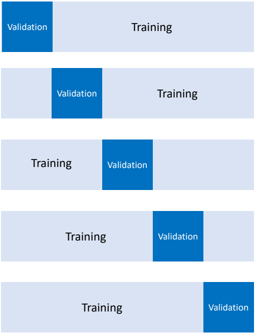

Figure 7: *Five-fold cross-validation.*

Cross-Validation

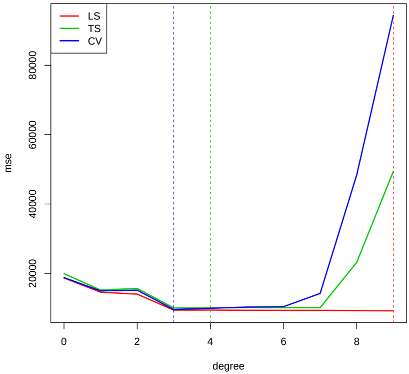

Figure 8: *MSE: Linear regression.* Learning set, test set, and cross-validated MSE for linear regression fits for polynomials of degree 0 to 9 ($n = 500$, $n^* = 10,000$). Dashed lines indicate risk minimizer.

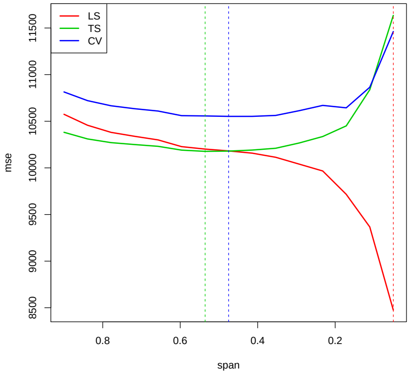

Figure 9: *MSE: Robust local regression.* Learning set, test set, and cross-validated MSE for loess fits for spans ranging from 0.05 to 0.90 ($n = 500$, $n^* = 10,000$). Dashed lines indicate risk minimizer.