Resampling

Machine Learning
Prof. Neylson Crepalde

Some of the figures in this presentation are taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani

Resampling, ou reamostragem, é um conjunto de técnicas fundamental nas análises estatísticas modernas. Elas consistem de retirar repetidamente, de maneira aleatória, várias amostras do *dataset* sob investigação para obter maiores informações sobre os modelos estimados.

Por exemplo, para estimar a variabilidade do ajuste de uma regressão linear, podemos tirar várias amostras dos dados de treino, ajustar um novo modelo para cada nova amostra e então verificar em que medida os resultados do modelo diferem entre as amostras.

Iremos, nesta unidade, discutir dois dos procedimentos de *resampling* mais utilizados, o **cross-validation** e o **bootstrap**.

Cross-validation

A divisão em *train set* e *test set*

Na unidade anterior, discutimos a distinção entre o erro dos dados de treino e o erro dos dados de teste. O erro dos dados de teste é o erro médio que resulta do uso de um método de *machine learning* para predizer a resposta de novas observações, isto é, uma medida que não foi usada no treino do modelo. Damos um modelo de ML como garantido se seus resultados possuem um erro de teste baixo. O erro de teste pode ser facilmente calculado se um *test set* está disponível. Entretanto, infelizmente na grande maioria das vezes não é esse o caso. Em contrapartida, o erro de treino pode ser facilmente calculado aplicando um modelo aos dados de treino. Contudo, o erro de treino é normalmente bem diferente do erro de teste e o primeiro pode subestimar dramaticamente o segundo.

Validation set

Para contornar o problema da falta de dados de teste, podemos criando uma subamostra que funcione como **set de validação**. Nessa abordagem, tiramos uma amostra aleatória de nossos dados para estimarmos o erro de teste de nossos modelos.

FIGURE 5.1. *A schematic display of the validation set approach. A set of n observations are randomly split into a training set (shown in blue, containing observations 7, 22, and 13, among others) and a validation set (shown in beige, and containing observation 91, among others). The statistical learning method is fit on the training set, and its performance is evaluated on the validation set.*

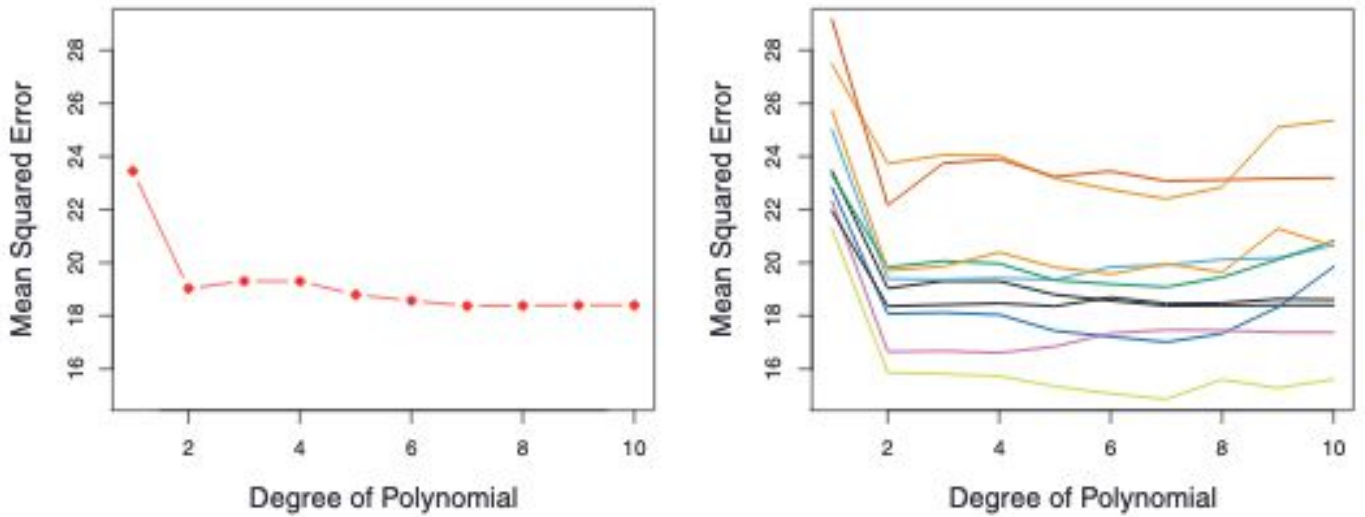

FIGURE 5.2. *The validation set approach was used on the* `Auto` *data set in order to estimate the test error that results from predicting* `mpg` *using polynomial functions of* `horsepower`. Left: *Validation error estimates for a single split into training and validation data sets.* Right: *The validation method was repeated ten times, each time using a different random split of the observations into a training set and a validation set. This illustrates the variability in the estimated test MSE that results from this approach.*

Leave-one-out cross-validation

No *Leave-one-out cross-validation* (LOOCV) ao invés de extrair uma subamostra dos dados para servir como set de validação, apenas um caso é usado para validação e o modelo é treinado em um subset de tamanho *n-1*.

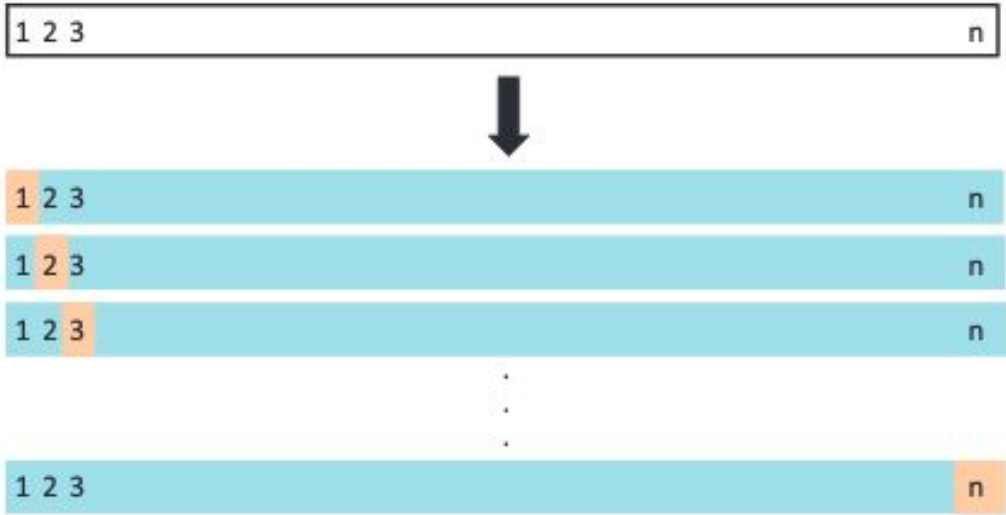

FIGURE 5.3. *A schematic display of LOOCV. A set of n data points is repeatedly split into a training set (shown in blue) containing all but one observation, and a validation set that contains only that observation (shown in beige). The test error is then estimated by averaging the n resulting MSE's. The first training set contains all but observation 1, the second training set contains all but observation 2, and so forth.*

LOOCV portanto é um método que provê um estimador não viesado para MSE (medida de erro para regressão). Entretanto, é um estimador muito pobre pois apenas um caso é utilizado. Podemos repetir esse procedimento até que obtenhamos n estimadores de MSE. LOOCV estimador para MSE seria a média de todos os MSE estimados.

Pelo fato de LOOCV fazer com que o modelo seja implementado n vezes, ele é computacionalmente muito custoso. Nos casos de regressão linear ou polinomial, James et. al (2013) apresentam uma fórmula que reduz o custo à computação de um modelo:

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2$$

Embora a fórmula só se sustente para esse caso, LOOCV é um método bastante geral e funciona bem com regressão logística e LDA. Nesses casos, os modelos precisam ser ajustados n vezes.

K-Fold cross-validation

K-Fold cross-validation é uma alternativa ao LOOCV bem menos custosa computacionalmente. Consiste em dividirmos o banco em *k* partes e utilizarmos uma para validação e *k - 1* para treino. O processo é repetido k vezes. O MSE final será a média de k MSE's.

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

Ciências de Dados

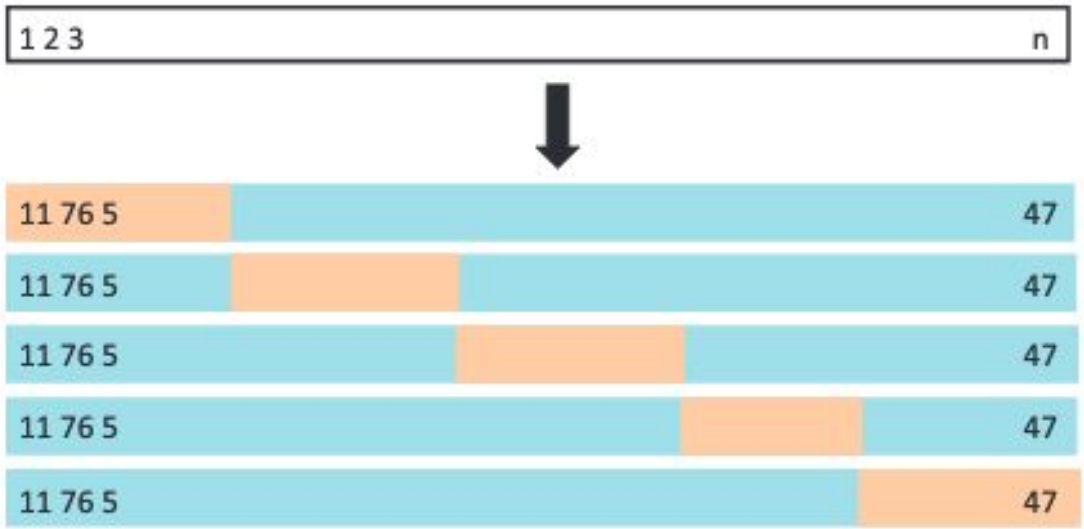

FIGURE 5.5. *A schematic display of 5-fold CV. A set of n observations is randomly split into five non-overlapping groups. Each of these fifths acts as a validation set (shown in beige), and the remainder as a training set (shown in blue). The test error is estimated by averaging the five resulting MSE estimates.*

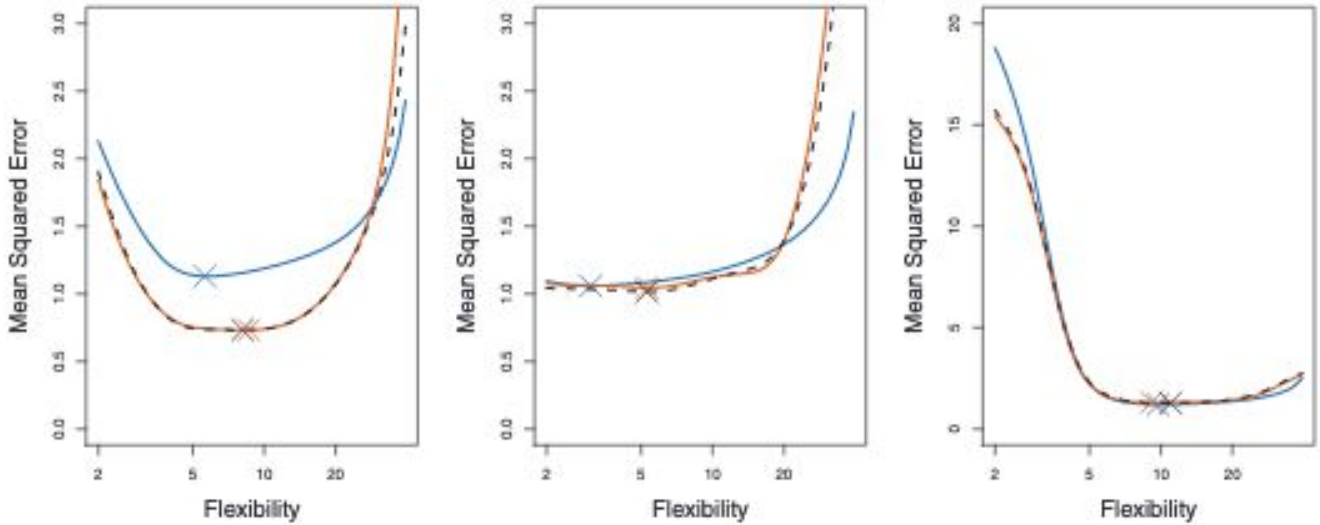

FIGURE 5.6. *True and estimated test MSE for the simulated data sets in Figures 2.9 (*left*), 2.10 (*center*), and 2.11 (*right*). The true test MSE is shown in blue, the LOOCV estimate is shown as a black dashed line, and the* 10*-fold CV estimate is shown in orange. The crosses indicate the minimum of each of the MSE curves.*

Cross-validation para classificação

Até agora tratamos do uso de *cross-validation* para modelos de regressão. O procedimento adotado para modelos de classificação é exatamente o mesmo excetuando-se a medida de ajuste escolhida. Nos casos de classificação, podemos trabalhar com a quantidade de observações classificadas errado. Nesse caso, o erro LOOCV pode ser definido por:

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{Err}_i,$$

Exemplo:

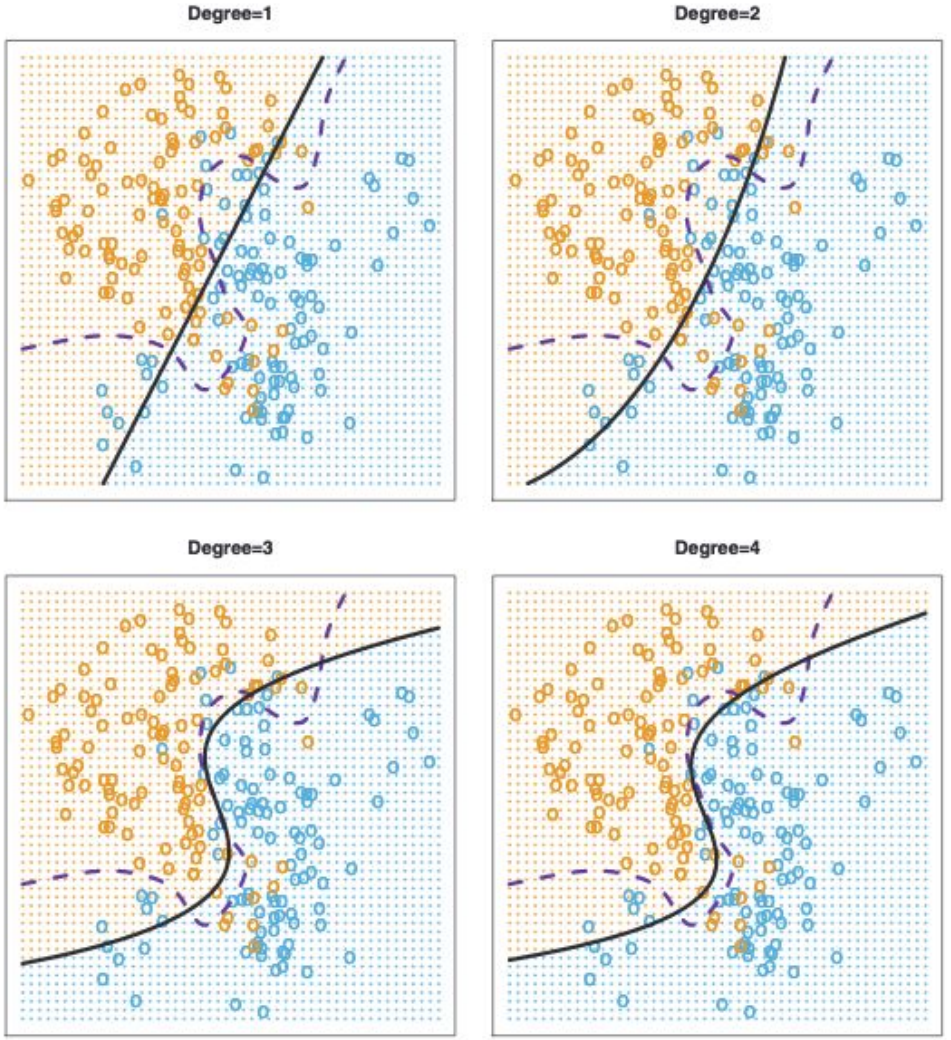

FIGURE 5.7. *Logistic regression fits on the two-dimensional classification data displayed in Figure 2.13. The Bayes decision boundary is represented using a purple dashed line. Estimated decision boundaries from linear, quadratic, cubic and quartic (degrees 1–4) logistic regressions are displayed in black. The test error rates for the four logistic regression fits are respectively* 0.201, 0.197, 0.160, *and* 0.162, *while the Bayes error rate is* 0.133.

Ciências de Dados

No mundo real, não sabemos qual é o erro verdadeiro do modelo logístico polinomial. Podemos tentar aproximá-lo utilizando cross-validation.

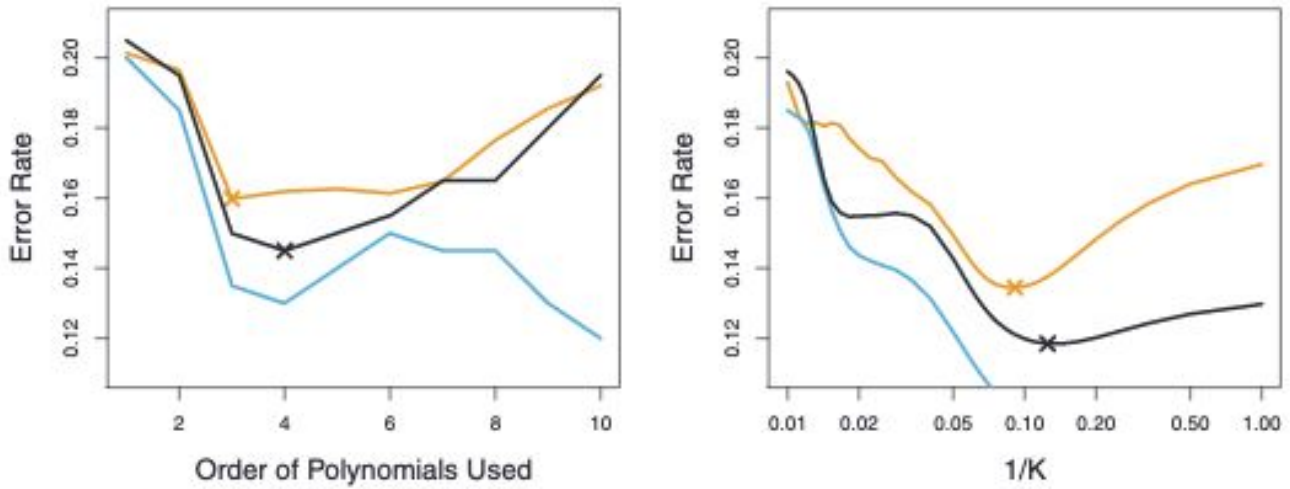

FIGURE 5.8. *Test error (brown), training error (blue), and 10-fold CV error (black) on the two-dimensional classification data displayed in Figure 5.7.* Left: *Logistic regression using polynomial functions of the predictors. The order of the polynomials used is displayed on the x-axis.* Right: *The KNN classifier with different values of K, the number of neighbors used in the KNN classifier.*

Ciências de Dados

Exercício

Bootstrap

Bootstrap

O Bootstrap é uma ferramenta estatística amplamente aplicável e extremamente poderosa que pode ser usada para quantificar a incerteza associada a um dado estimador ou um método de aprendizado estatístico.

Ela funciona obtendo *datasets* distintos retirando repetidas observações do *dataset* original. As observações são retiradas aleatoriamente **com reposição**, isto é, com a possibilidade de que alguns casos apareçam mais de uma vez. Com esse método podemos estimar o erro padrão de estatísticas de interesse ou o intervalo de confiança de estimadores, por exemplo.

O procedimento é ilustrado na figura a seguir.

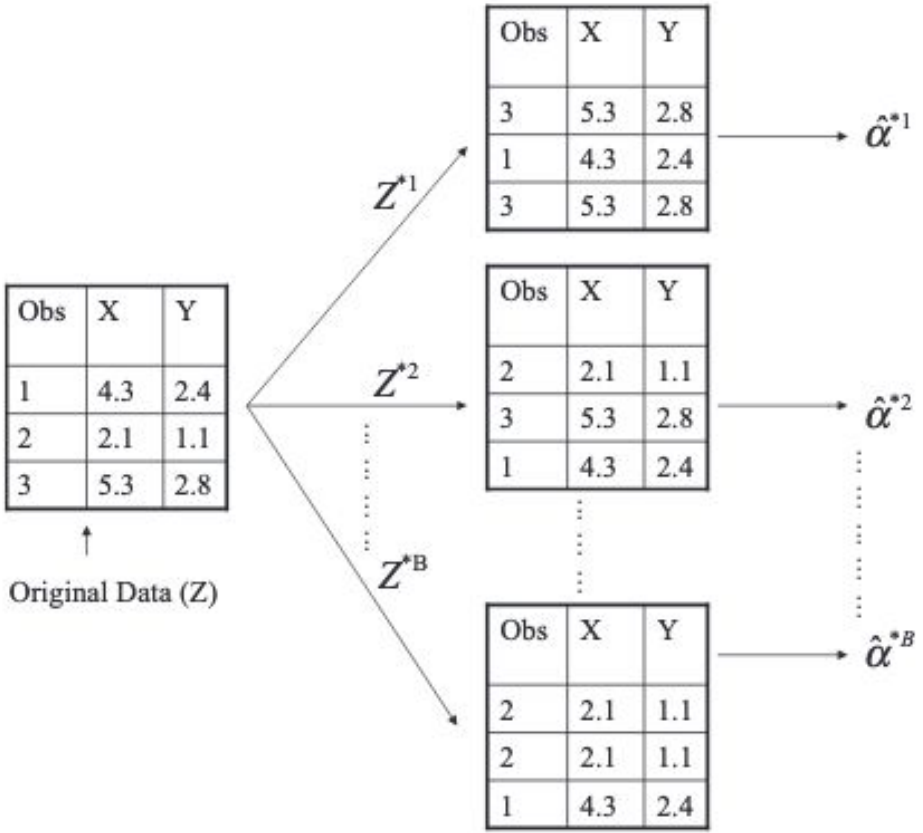

FIGURE 5.11. *A graphical illustration of the bootstrap approach on a small sample containing $n = 3$ observations. Each bootstrap data set contains n observations, sampled with replacement from the original data set. Each bootstrap data set is used to obtain an estimate of α.*

Atividade Avaliativa!!

Ciências de Dados