Mineração de Dados / Statistical Machine Learning

Lista 1

Rafael Izbicki

Lista em até trios. Apenas um do grupo deve entregar a lista. Por favor entregue a lista em PDF.

NÃO COPIE!

Exercício 1. Considere o banco `houses_to_rent_v2` (https://www.dropbox.com/s/8auhk2zaogovjvn/houses_to_rent_v2.csv?dl=0), que contém o valor (em reais) do aluguel de imóveis no Brasil. Você pode trabalhar apenas com os imóveis localizados em São Paulo, Rio de Janeiro e Belo Horizonte.

(a) Divida o conjunto de dados em para treinamento e teste. Explique como decidiu qual porcentagem deixar para cada um.

(b) Utilizando o conjunto de treinamento, ajuste uma regressão (i) via mínimos quadrados, (ii) via lasso (usando validação-cruzada no treinamento para escolher λ) e *(iii) (Pós-graduação apenas) regressão ridge*. Qual o melhor valor de λ encontrado para o lasso?

(c) Qual dos métodos acima apresentou melhores resultados? Responda essa pergunta utilizando o conjunto de teste e o melhor valor de λ encontrado. Inclua os intervalos de confiança para o risco preditivo nos seus resultados.

(d) Interprete os resultados do melhor modelo encontrado (via coeficientes). Ele faz sentido?

(e) Includa todas as interações entre as variáveis observadas e repita o ajuste do método de mínimos quadrados e do lasso. Como esses ajustes se comparam em relação aos anteriores? Qual foi o melhor modelo encontrado? Esses resultados são esperados?

Exercício 2 (Pós-graduação)**.** Mostre que

$$\mathbb{E}\left[(Y - g(\mathbf{X}))^2 \middle| \mathbf{X} = \mathbf{x}\right] = \mathbb{V}[Y|\mathbf{X} = \mathbf{x}] + (r(\mathbf{x}) - \mathbb{E}[g(\mathbf{x})])^2 + \mathbb{V}[g(\mathbf{x})].$$

Exercício 3 (Pós-graduação)**.** Seja $0 < \alpha < 1$ fixo e considere a função de perda

$$L(g; (\mathbf{X}, Y)) = (g(\mathbf{X}) - Y)(\mathbb{I}(Y \leq g(\mathbf{X})) - \alpha).$$

Qual a função g que minimiza a função de risco (aleatório apenas em $(\mathbf{X}, Y)$) correspondente? Interprete e justifique.