Mineração de Dados

Lista 1

Rafael Izbicki

Lista em trios.

NÃO COPIE!

Exercício 1. Baixe os dados `worldDevelopmentIndicators.csv` em http://www.rizbicki.ufscar.br/dados/worldDevelopmentIndicators.csv, que contém os dados do PIB per capita (X) e a expectativa de vida (Y) de diversos países. O objetivo é criar preditores de Y com base em X. Em aula vimos como isso pode ser feito através de polinômios. Aqui, faremos isso via expansões de Fourier.

1. Normalize a covariável de modo que $x \in (0,1)$. Para isso, faça $x = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$, onde $x_{\min}$ e $x_{\max}$ são os valores mínimos e máximos de x segundo a amostra usada.

2. Usando o método dos mínimos quadrados e a validação cruzada do tipo *leave-one-out*, estime o risco das regressões

   $$g(x) = \widehat{\beta}_0 + \widehat{\beta}_1 \sin(2\pi x) + \widehat{\beta}_2 \cos(2\pi x) + \widehat{\beta}_3 \sin(2\pi 2x) + \widehat{\beta}_4 \cos(2\pi 2x) + \ldots + \widehat{\beta}_{2p-1} \sin(2\pi p x) + \widehat{\beta}_{2p} \cos(2\pi p x)$$

   para $p = 1, \ldots, 30$.

3. Plote o gráfico do risco estimado vs p. Qual o valor de p escolhido? Denotaremos ele por p_{esc}

4. Plote os das curvas ajustadas para $p = 1$, $p = p_{esc}$ e $p = 30$ (com um ajuste usando a amostra toda) sob o gráfico de dispersão de X por Y. Qual curva parece mais razoável? Use um grid de valores entre 0 e 1 para isso. Como estes ajustes se comparam com o visto em aula via polinômios? Discuta.

5. Plote o gráfico de valores preditos versus observados para $p = 1$, $p = p_{esc}$ e $p = 30$ (não se esqueça de usar o *leave-one-out* para calcular os valores preditos! Caso contrário você terá problemas de overfitting novamente). Qual p parece ser o mais razoável?

6. Quais vantagens e desvantagens de se usar validação cruzada do tipo *leave-one-out* versus o *data-splitting*?

Dica: caso você use a função `lm` do R, considere usar a função `paste` para automatizar as fórmulas da regressão. Não faça o ajuste para cada p na mão!