Aprendizaod de Máquina

Lista 1 - Entrega 29/08

Marcos Prates

Lista em duplas escolhidas pelo professor. Se você usou R ou outro programa orientado a programação para resolver os exercícios práticos, submeta junto com as respostas o código utilizado para resolvê-las **(o código pode ser por email, as respostas devem ser impressas)**. Se você usou algum programa visual, explique qual programa foi usado, e como o problema foi resolvido nele (que funcionalidades foram usadas etc).

NÃO COPIE!

Exercício 1. Baixe no site os 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. **[0.5 pts]** 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. **[5.0 pts]** Usando o método dos mínimos quadrados e a validação cruzada do tipo *leave-one-out*, estime o erro quadrático médio 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. **[1.0 pts]** Plote o gráfico do risco estimado vs p. Qual o valor de p escolhido? Denotaremos ele por p_{esc}

4. **[1.5 pts]** Plote os das curvas ajustadas para $p = 1$, $p = p_{esc}$ e $p = 30$ 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. **[1.5 pts]** Plote o gráfico de valores preditos versus ajustados 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. **[0.5 pts]** Quais vantagens e desvantagens de se usar validação cruzada do tipo *leave-one-out* versus o *data-splitting*?

7. **[0.5 pts]** Ajuste a regressão Lasso (Frequentista e Bayesiana) e discuta os resultados encontrados.

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!

Exercício 2. Neste exercício você irá implementar algumas técnicas vistas em aula para o banco de dados das faces. O objetivo aqui é conseguir criar uma função que consiga predizer para onde uma pessoa está olhando com base em uma foto. Iremos aplicar o KNN para esses dados, assim como uma regressão linear. Como não é possível usar o método dos mínimos quadrados quando o número de covariáveis é maior que o número de observações, para esta segunda etapa iremos usar o lasso.

(a) **[1.0 pts]** Leia o banco `dadosFacesAltaResolucao.txt`. A primeira coluna deste banco contém a variável que indica a direção para a qual o indivíduo na imagem está olhando. As outras covariáveis contém os pixels relativos a essa imagem, que possui dimensão 64 por 64. Utilizando os comandos fornecidos, plote 5 imagens deste banco.

Divida o conjunto fornecido em treinamento (aproximadamente 60% das observações), validação (aproximadamente 20% das observações) e teste (aproximadamente 20% das observações). Utilizaremos o conjunto de treinamento e validação para ajustar os modelos. O conjunto de teste será utilizado para testar sua performance.

(b) **[0.5 pts]** Qual o número de observações? Qual o número de covariáveis? O que representa cada covariável?

(c) **[1.0 pts]** Para cada observação x do conjunto de teste, calcule o estimador da função de regressão $r(x)$ dado pelo método dos k vizinhos mais próximos com $k = 5$. Você pode usar as funções vistas em aula.

(d) **[1.0 pts]** Utilize validação cruzada (*data splitting*) para escolher o melhor k. Plote k vs Risco estimado.

(e) **[1.0 pts]** Utilizando o conjunto de teste, estime o risco do KNN para o melhor k. Plote os valores preditos versus os valores observados para o conjunto de teste. Inclua a reta identidade.

(f) **[2.0 pts]** Ajuste uma regressão linear para os dados usando o conjunto de treinamento mais o de validação via lasso (lembre-se que a função que ajusta o lasso no R já faz validação cruzada automaticamente: ao contrário do KNN, neste caso não é necessário separar os dados em treinamento e validação). Qual o lambda escolhido? Plote lambda vs Risco estimado.

(g) **[1.5 pts]** Utilizando o conjunto de teste, estime o risco do lasso para o melhor lambda. Plote os valores preditos versus os valores observados para o conjunto de teste. Inclua a reta identidade.

(h) **[1.0 pts]** Quantos coeficientes foram estimados como sendo zero?

(i) **[1.0 pts]** Qual modelo teve melhores resultados: regressão linear via lasso ou KNN?