Aprendizagem de Máquina I – Lista 01
Referente aos slides "01 Intro"

Prof. Hugo Carvalho

Fontes para os exercícios

- [ITSL] Gareth James, Daniela Witten, Trevor Hastie, Rob Tibshirani & Jonathan Taylor - *An Introduction to Statistical Learning, with Applications in Python* [baixe aqui]
- [AME] Rafael Izbicki & Tiago Mendonça dos Santos - *Aprendizado de Máquina: Uma Abordagem Estatística* [baixe aqui]

Questões do livro

- [ITSL] Capítulo 2
  - Questões conceituais: 1, 2, 3, 4, 5, 6
  - Questões aplicadas: 1

Questões avulsas

Questão 1: O objetivo dessa questão é tornar rigorosa e provar a equação do balanço entre viés e variância. Para isso, siga os passos indicados abaixo.

Obs.: Cristopher Bishop em "Pattern Recognition and Machine Learning", Sec. 3.2 (pág. 147) dá alguns vagos indícios de como fazer essa conta.

Obs^2.: Eu sei que essa questão é "emocionante". Me lembro que demorei um bom tempo até conseguir desmembrar a conta e chegar nesse passo-a-passo. Qualquer dificuldade, não hesitem em perguntar!

a) Assuma que a relação entre $Y \in \mathbb{R}$ e $\mathbf{X} \in \mathbb{R}^p$ é dada por $Y = f(\mathbf{X}) + \varepsilon$, onde a função $f : \mathbb{R}^p \to \mathbb{R}$ é desconhecida e ε é um erro de média zero, independente de $\mathbf{X}$. Denote por $\hat{f}$ uma estimativa de f e assuma que $\mathbf{X}$ e Y sigam uma distribuição conjunta cuja densidade é dada por $p(\mathbf{x}, y)$. Mostre que a "melhor" possível estimativa (no sentido de minimizar a função custo quadrática) é dada por $\hat{f}_{\text{opt}}(\mathbf{x}) = \mathbb{E}_Y[Y|\mathbf{X} = \mathbf{x}]$. Para isso, mostre que o seguinte valor esperado é mínimo quando $\hat{f} = \hat{f}_{\text{opt}}$:

$$\mathbb{E}_{\mathbf{X},Y}[(\hat{f}(\mathbf{X}) - Y)^2] = \iint (\hat{f}(\mathbf{x}) - y)^2 p(\mathbf{x}, y)\ d\mathbf{x}dy.$$

b) Seja agora $\hat{f}$ uma estimativa de f obtida a partir de um determinado conjunto de observações $\mathcal{D}$. Explicitemos tal fato agora escrevendo $\hat{f}(\mathbf{x})$ como $\hat{f}(\mathbf{x}; \mathcal{D})$. Fixe $\mathbf{x}$ e considere a distância quadrática entre $\hat{y} = \hat{f}(\mathbf{x}; \mathcal{D})$ e sua "melhor" previsão possível, ou seja, $(\hat{f}(\mathbf{x}; \mathcal{D}) - \hat{f}_{\text{opt}}(\mathbf{x}))^2$. Mostre que o valor esperado de tal quantidade com respeito a todos os possíveis conjuntos de dados é dada por

$$\mathbb{E}_{\mathcal{D}}[(\hat{f}(\mathbf{x}; \mathcal{D}) - \hat{f}_{\text{opt}}(\mathbf{x}))^2] = (\mathbb{E}_{\mathcal{D}}[\hat{f}(\mathbf{x}; \mathcal{D})] - \hat{f}_{\text{opt}}(\mathbf{x}))^2 - \mathbb{E}_{\mathcal{D}}[(\hat{f}(\mathbf{x}; \mathcal{D}) - \mathbb{E}_{\mathcal{D}}[\hat{f}(\mathbf{x}; \mathcal{D})])^2].$$

Dica: $0 = \mathbb{E}_{\mathcal{D}}[\hat{f}(\mathbf{x};\mathcal{D})] - \mathbb{E}_{\mathcal{D}}[\hat{f}(\mathbf{x};\mathcal{D})]$.

Finalmente, argumente que o primeiro desses termos é o que chamamos de "viés" (ao quadrado) e o segundo de "variância".

c) Note que o resultado obtido acima é quase o que queremos, a menos do termo sobre a variância do erro de observação. De modo a incorporá-lo, retorne ao item a) e mostre que $\mathbb{E}_{\mathbf{X},Y}[(\hat{f}(\mathbf{X}) - Y)^2]$ pode ser escrita como

$$\mathbb{E}_{\mathbf{X},Y}[(\hat{f}(\mathbf{X}) - Y)^2] = \int (\hat{f}(\mathbf{x}) - \hat{f}_{\text{opt}}(\mathbf{x}))^2 p(\mathbf{x})\ d\mathbf{x} + \iint (\hat{f}_{\text{opt}}(\mathbf{x}) - y)^2 p(\mathbf{x}, y)\ d\mathbf{x}dy.$$

Dica: Novamente, 0 = alguma coisa − a mesma coisa. Descubra quem é essa coisa.

d) Juntando os resultados dos itens b) e c), conclua que

$$\begin{aligned}\mathbb{E}_{\mathbf{X},Y,\mathcal{D}}[(\hat{f}(\mathbf{X};\mathcal{D}) - Y)^2] = &\int (\mathbb{E}_{\mathcal{D}}[\hat{f}(\mathbf{x};\mathcal{D})] - \hat{f}_{\text{opt}}(\mathbf{x}))^2 p(\mathbf{x})\ d\mathbf{x} + \\ &\int \mathbb{E}_{\mathcal{D}}[(\hat{f}(\mathbf{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}[\hat{f}(\mathbf{x};\mathcal{D})])^2] p(\mathbf{x})\ d\mathbf{x} + \\ &\iint (\hat{f}_{\text{opt}}(\mathbf{x}) - y)^2 p(\mathbf{x}, y)\ d\mathbf{x}dy.\end{aligned}$$

Para concluir, argumente que o primeiro desses termos é o que chamamos de "viés" (ao quadrado), segundo é a "variância" e o terceiro está relacionado com o ruído inerente das observações.