Aprendizagem de Máquina I – Lista 03
Referente aos slides "04 Metodos nao-parametricos (KNN)" e "05 Metodos nao-parametricos (KNN) - aspectos teoricos"

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 3
  - Leia a seção "3.5 Comparison of Linear Regression with K-Nearest Neighbors".

Questões avulsas

Questão 1: O objetivo dessa questão é estudar um pouco da *maldição da dimensionalidade*. De modo mais concreto, queremos estudar a seguinte afirmação: "à medida que p cresce, o volume do hipercubo $\mathbb{H}^p = [0,1]^p$ fica cada vez mais concentrado em sua 'casca' do que em seu interior." Para isso, faça o que se pede abaixo.

a) a) Seja $\mathbb{H}^p_\varepsilon$ o hipercubo contido em $\mathbb{H}^p$ obtido removendo-se uma "gordurinha" de tamanho ε de sua beirada, ou seja, $\mathbb{H}^p_\varepsilon = [\varepsilon/2, 1-\varepsilon/2]^p$. Mostre que o volume de $\mathbb{H}^p_\varepsilon$ converge para zero, para todo $\varepsilon > 0$, quando $p \to \infty$.

b) b) Fixe algum valor de $0 < \varepsilon < 1$ à sua escolha e simule n vetores de tamanho p de observações de uma variável aleatória uniforme no intervalo $[0,1]$. Qual é a proporção desses vetores que estão dentro do hipercubo $\mathbb{H}^p_\varepsilon$? Faça uma explicação, baseada em tal experimento, para justificar empiricamente o resultado que você provou no item a). Repare que se p é suficientemente grande, será muito raro que qualquer desses vetores não tenha ao menos uma entrada menor que $\varepsilon/2$ ou maior que $1-\varepsilon/2$.

Questão 2: O objetivo dessa questão é provar o que foi discutido sobre a maldição da dimensionalidade ao final dos slides "04 Metodos nao-parametricos (KNN)". Considere um conjunto de dados artificial de n observações, onde cada observação consiste de um vetor de p atributos, sendo cada entrada uniformemente distribuída no intervalo $(0,1)$. Sejam $\mathbf{X}^{(i)}$ e $\mathbf{X}^{(j)}$ dois vetores independentes de observações, gerados de acordo com esse processo, ou seja, ambos de tamanho p, consistindo de entradas independentes e

identicamente distribuídas com distribuição uniforme no intervalo $(0, 1)$. Estude o que acontece com a quantidade abaixo, chamada de *coeficiente de variação* à medida que $p \to \infty$:

$$\text{cv}\left(\left\|\mathbf{X}^{(i)} - \mathbf{X}^{(j)}\right\|^2\right) = \frac{\sqrt{\mathbb{V}\left(\left\|\mathbf{X}^{(i)} - \mathbf{X}^{(j)}\right\|^2\right)}}{\mathbb{E}\left[\left\|\mathbf{X}^{(i)} - \mathbf{X}^{(j)}\right\|^2\right]}.$$

Relacione esse resultado com o comportamento apresentado nos slides "04 Metodos nao-parametricos (KNN)".