Aprendizagem de Máquina I – Lista 05
Referente aos slides "08 Classificacao e classificadores gaussianos" e "09 Metricas para classificacao"

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 4
  - Questões conceituais: 7, 8

Questões avulsas

Questão 1: Prove o resultado a seguir para um problema de classificação multiclasse, ou seja, quando $\mathcal{C}$ tem mais de dois elementos (lembre-se que fizemos, em aula, a prova para o caso de classificação binária):

Teorema 1. *A função $g : \mathbb{R}^p \to \mathcal{C}$ que minimiza o risco $R(g) = \mathbb{E}[\mathbb{I}(Y \neq g(\mathbf{X}))] = \mathbb{P}(Y \neq g(\mathbf{X}))$ é o classificador de Bayes, dado por $g(\mathbf{x}) = \text{argmax}_{d \in \mathcal{C}} \mathbb{P}(Y = d | \mathbf{X} = \mathbf{x})$.*

Questão 2: Agora vamos provar um resultado análogo no caso de classificação binária ($\mathcal{C} = \{0, 1\}$), mas assumindo que cada erro tem um peso diferente. Formalmente, encontre o classificador $g : \mathbb{R}^p \to \mathcal{C}$ que minimiza a função risco abaixo:

$$\begin{aligned} R(g) &= \mathbb{E}[\ell_1 \mathbb{I}(Y \neq g(\mathbf{X}) \text{ e } Y = 0) + \ell_0 \mathbb{I}(Y \neq g(\mathbf{X}) \text{ e } Y = 1)] \\ &= \ell_1 \mathbb{P}(Y \neq g(\mathbf{X}) \text{ e } Y = 0) + \ell_0 \mathbb{P}(Y \neq g(\mathbf{X}) \text{ e } Y = 1). \end{aligned}$$

Questão 3: Vimos em aula métricas para avaliar o resultado de uma classificação no caso binário. Estude generalizações dessas métricas para o caso multiclasse. Para isso você precisará estudar duas formas de generalizar classificadores binários para o caso multiclasse, chamadas de *one-versus-one* (OvO) e *one-versus-all* (OvA). Alguns classificadores são naturalmente generalizáveis para o caso multiclasse (por exemplo, classificadores Naive Bayes), mas isso não é o caso para outros classificadores que veremos ao longo do curso, por isso essas generalizações são importantes.

Questão 4: Desde a primeira parte curso, referente a modelos regressão, implementamos o seguinte procedimento na validação cruzada: utilizamos validação cruzada no conjunto de treinamento para encontrar o "melhor" modelo dentro de cada classe (por exemplo, o "melhor" λ_R para Ridge, o "melhor" λ_L para o Lasso, o "melhor" k para o KNN, etc.), retreinamos esses "melhores" modelos de cada classe no conjunto de treinamento todo e escolhemos pelo "melhorzão" comparando o desempenho deles no conjunto de teste. No caso de regressão não conseguimos provar rigorosamente que isso é bom, devido à natureza contínua (e potencialmente ilimitada) da função risco R. Porém, no caso discreto, como o risco é uma variável aleatória limitada, conseguimos provar coisas interessantes. Para isso, precisaremos de um resultado preliminar:

Teorema 2 (Desigualdade de Hoeffding)**.** *Para $W_1, \ldots, W_n$ variáveis aleatórias independentes onde cada W_i tem suporte em (a_i, b_i), para todo $\varepsilon > 0$ vale que:*

$$\mathbb{P}(|\overline{W}_n - \mathbb{E}[\overline{W}_n]| > \varepsilon) \leq 2e^{-2n^2\varepsilon^2 / \sum_{i=1}^n (b_i - a_i)^2},$$

onde $\overline{W}_n = \frac{1}{n}\sum_{i=1}^n W_i$.

O resultado que queremos provar pode ser formalmente enunciado como:

Teorema 3. *Seja $\mathbb{G} = \{g_1, \ldots, g_N\}$ uma classe de classificadores estimados com base em um conjunto de treinamento e seja $\widehat{R}(g)$ o erro estimado de g com base no conjunto de teste (de tamanho s):*

$$\widehat{R}(g) = \frac{1}{s}\sum_{i=1}^{s} \mathbb{I}(Y_i \neq g(\mathbf{X}_i)) \approx R(g).$$

Seja g^* o modelo que minimiza o risco real $R(g)$, dentre $g \in \mathbb{G}$, e seja $\widehat{g}$ o modelo que minimiza o risco estimado $\widehat{R}(g)$, dentre $g \in \mathbb{G}$. Então, com probabilidade de no máximo ε, vale que:

$$|R(\widehat{g}) - R(g^*)| > 2\sqrt{\frac{1}{2s}\log\frac{2N}{\varepsilon}}.$$

Para provar esse resultado siga o roteiro abaixo.

a) Use a desigualdade de Hoeffding para limitar superiormente a quantidade $\mathbb{P}(|\widehat{R}(g) - R(g)| > \delta)$, fixado um $\delta > 0$ e um $g \in \mathbb{G}$.

b) Argumente que

$$\mathbb{P}\left(\max_{g\in\mathbb{G}} \left|\widehat{R}(g) - R(g)\right| > \delta\right) = \mathbb{P}\left(\bigcup_{g\in\mathbb{G}} \left\{\left|\widehat{R}(g) - R(g)\right| > \delta\right\}\right),$$

e use que $\mathbb{P}(\cup A_i) \leq \sum \mathbb{P}(A_i)$ juntamente com o resultado obtido no item a) para limitar superiormente o lado esquerdo da expressão acima.

c) Escolha um δ adequado para concluir o resultado desejado. Tome cuidado pois o resultado desejado envolve $R(g^*)$ e o que provamos acima não tem essa quantidade. Você terá que fazer uma malangradem para conseguir fazer aparecer o $R(g^*)$.

d) Interprete o resultado obtido, relacionando-o com a interpretação proposta no início da presente questão.

e) À luz do resultado provado e de sua interpretação, argumente porque a estratégia de "treinar todo mundo no conjunto de treino e compará-los no conjunto de teste para encontrar o 'melhor' modelo" é, provavelmente, pior do que a estratégia proposta no início da questão.

f) Justifique porque essa prova não vale para um problema de regressão.

g) Prove a desigualdade de Hoeffding.