Lista de Exercícios - Classificação e Teoremas Limite

Renato Assunção - DCC, UFMG

2018

1. Replicar a análise de classificação usando a função LDA de Fisher em duas páginas da web (uma sendo a sequência da seguinte): `http://www.aaronschlegel.com/discriminant-analysis/` e `https://www.r-bloggers.com/classification-with-linear-discriminant-analysis/`. Os dados não estão imediatamente visíveis apontado pelas páginas mas eu os coloquei na página da nossa disciplina.

2. Existem duas classes ou populações, 1 e 2, preentes nas proporções positivas π_1 e π_2 com $\pi_1+\pi_2 = 1$. Suponha que o vetor aleatório contínuo $\mathbf{X} = (X_1, \ldots, X_p)$ com p variáveis possua as densidades $f_1(\mathbf{x})$ e $f_2(\mathbf{x})$ quando o indivíduo pertence à população 1 ou 2, respectivamente. Sejam $c(1|2)$ o custo do erro de classificar erradamente no grupo 1 um indivíduo que seja do grupo 2. Analogamente, defina o custo do outro erro $c(2|1)$. A região ótima R_1 de classificação no grupo 1 é dada pela seguinte região do espaço $\mathbb{R}^p$:

$$R_1 = \left\{ \mathbf{x} \in \mathbb{R}^p \text{ tais que } \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} > \frac{c(1|2)}{c(2|1)} \frac{\pi_2}{\pi_1} \right\}$$

   Note que a ordem das populações na última fração é oposta à ordem na razão das densidades. Isto é, comparamos f_1/f_2 com π_2/π_1.

   - Suponha que $c(1|2) = c(2|1)$ e que $\pi_1 = \pi_2$. Neste caso, a regra ótima fica reduzida a uma simples comparação. Qual é esta regra de classificação?
   - Imagine agora que $\pi_1 = 0.01$ e que $c(1|2) = c(2|1)$. Para tornar as coisas mais concretas, suponha que a população 1 sejam portadores de certo vírus e a população 2, os demais. A regra simples do item acima fica modificada. Agora não basta que $f_1(\mathbf{x})$ seja maior que $f_2(\mathbf{x})$. Ela precia ser bem maior que $f_2(\mathbf{x})$. Quantas vezes maior $f_1(\mathbf{x})$ deve ser para que classifiquemos o item com característica $\mathbf{x}$ em 1?
   - Suponha que os custos de má-classificação sejam muito diferentes. O custo de classificar o portador do vírus como são pode custar-lhe a vida ou a vida de outras pessoas. Por outro lado, o indivíduo saudável ser classificado como infectado custa mais exames confirmatórios, algumas medidas de isolamento e outras coisas que são relativamente menos custosas. Suponha que $c(1|2)$ seja 10 vezes menor que $c(2|1)$. Neste caso, com $\pi_1 = 0.01$, como a regra do item acima fica modificada?

3. Um programa é usado para classificar fotos de gatos (população 1) versus fotos de não-gatos (população 2). As fotos da população 1 (fotos de gatos) são chamadas de *relevantes*. O classificador seleciona algumas fotos para classificar no grupo 1 baseado em features aleatórias no vetor $\mathbf{X}$. A regra de classificação é representada pela função binária $D(\mathbf{X})$ que assume os valores 1 ou 2 dependendo do vetor aleatório $\mathbf{X}$ cair ou não na região R_1 de classificação no grupo 1.

   Haverá erros nesta classificação e queremos torná-los pequenos. Duas métricas muito populares para avaliar a qualidade de um classificador são: *precisão* (precision, em inglês) e *revocação* (recall, em inglês). A palavra revocação não é muito usada na linguagem diária. Ela siginifica “fazer voltar, retornar, chamar novamente”. Pode significar também revogação, anulamento de um contrato mas não é este o significado relevante para nosso contexto.

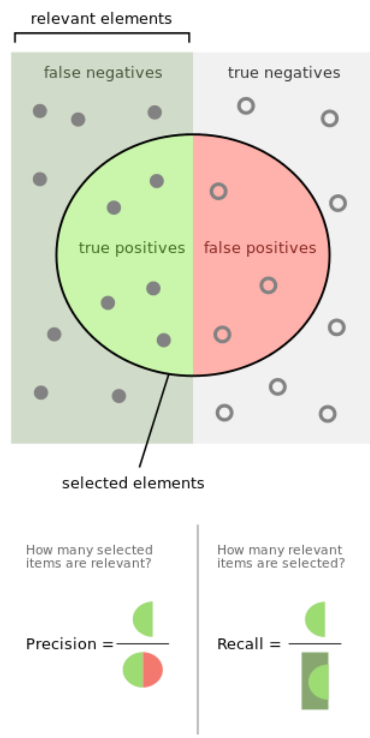

Figura 1: Retirado da Wikipedia.

- Precisão: $\mathbb{P}(\text{foto} \in \text{gatos} \mid \text{classificado como gato}) = \mathbb{P}(\mathbf{X} \in 1 | D(\mathbf{x}) = 1)$
- Revocação: $\mathbb{P}(\text{classificado como gato} | \text{foto} \in \text{gatos}) = \mathbb{P}(D(\mathbf{x}) = 1 | \mathbf{X} \in 1)$

É claro que, tanto para precisão quanto para revocação, quanto maior, melhor. Precisão e revocação são probabilidades condicionais usando os mesmos eventos A e B mas um deles é $\mathbb{P}(A|B)$ enquanto o outro é simplemente $\mathbb{P}(B|A)$. Sabemos que estas probabilidades podem ser muito diferentes. A Figura 1, retirada da página `Precision_and_recall` na Wikipedia, mostra itens nas suas classes reais: relevante (pop 1) ou não (pop 2). Mostra também a sua classificação na classe 1 (os itens dentro da elipse central) ou na classe 2 (os restantes). A Figura ainda mostra as probabilidades precisão e revocação como diagramas de Venn dos eventos envolvidos.

Marque V ou F nas afirmativas a seguir:

- A precisão mede o quanto os resultados da classificação são úteis.
- A revocação mede o quanto os resultados da aplicação da regra de classificação são completos.
- A soma de precisão e revocação é igual a 1.
- Precisão = Revocação $\times \dfrac{\mathbb{P}(\mathbf{X} \in 1)}{\mathbb{P}(D(\mathbf{x}) = 1)}$.
- Existe um trade-off entre precisão e revocação: se aumentarmos uma métrica, a outra tem de diminuir.

4. Existem duas classes ou populações, 1 e 2, preentes nas proporções positivas π_1 e π_2 com $\pi_1 + \pi_2 = 1$. Suponha que $pi_1 \approx 0$. O vetor aleatório contínuo $\mathbf{X} = (X_1, \ldots, X_p)$ com p variáveis possua as densidades $f_1(\mathbf{x})$ e $f_2(\mathbf{x})$ quando o indivíduo pertence à população 1 ou 2, respectivamente. Sejam $c(1|2)$ o custo do erro de classificar erradamente no grupo 1 um indivíduo que seja do grupo 2. Analogamente, defina o custo do outro erro $c(2|1)$. A regra de classificação é representada pela

função binária $D(\mathbf{X})$ que assume os valores 1 ou 2 dependendo do vetor aleatório $\mathbf{X}$ cair ou não na região R_1 de classificação no grupo 1.

- Uma regra de decisão que vai errar pouco será atribuir a classe 2 a todo e qualquer item: $D(\mathbf{X}) \equiv 2$ para todo valor de $\mathbf{X}$. Obtenha a probabilidade de classificação errada. A probabilidade é próxima de zero?
- Se o custo de má-clasificação for também desbalanceado, com $c(2|1) >> c(1|2)$, a estratégia anterior pode ser muito ruim. Obtenha o custo esperado de má-classificação (ECM) da regra anterior.

5. Você quer selecionar uma amostra para estimar a porcentagem θ de pessoas que vai votar num candidato X. Imagine que a resposta é uma v.a. X de Bernoulli com valores 1 e 0 (vai e não vai votar, respectivamente) e a probabilidade de sucesso é θ. As respostas de n indivíduos serão $X_1, X_2, \ldots, X_n$ e você vai estimar θ usando $\hat{\theta} = (X_1 + \ldots + X_n)/n$, a proporção amostral. Se você asumir que as respostas são variáveis aleatórias i.i.d., determine o tamanho n da amostra necessário para que o erro de estimação $|\hat{\theta} - \theta|$ seja menor que 0.02 com probabilidade 0.99. Para isto, assuma que você sabe que seu candidato está estacionado entre 15% e 35% dos eleitores (baseado em outras pesquisas mais antigas).

6. No problema acima, determine um intervalo da forma $I = (\hat{\theta} - c, \hat{\theta} + c)$ tal que a probabilidade $\mathbb{P}(\hat{\theta} - c \leq \theta \leq \hat{\theta} + c)$ seja aproximadamente igual ou maior que 0.95.