

Lista de Exercícios - Classificação e Teoremas Limite

Renato Assunção - DCC, UFMG

2018

1. Existem duas classes ou populações, 1 e 2, presentes nas proporções positivas π_1 e π_2 com $\pi_1 + \pi_2 = 1$. Suponha que o vetor aleatório contínuo $\mathbf{X} = (X_1, \dots, X_p)$ 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 :

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

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 precisa 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 não 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?
2. 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 significa “fazer voltar, retornar, chamar novamente”. Pode significar também revogação, anulamento de um contrato mas não é este o significado relevante para nosso contexto.

- Precisão: $\mathbb{P}(\text{foto} \in \text{gatos} \mid \text{classificado como gato}) = \mathbb{P}(\mathbf{X} \in 1 \mid D(\mathbf{x}) = 2)$
- Revocação: $\mathbb{P}(\text{classificado como gato} \mid \text{foto} \in \text{gatos}) = \mathbb{P}(D(\mathbf{x}) = 2 \mid \mathbf{X} \in 1)$

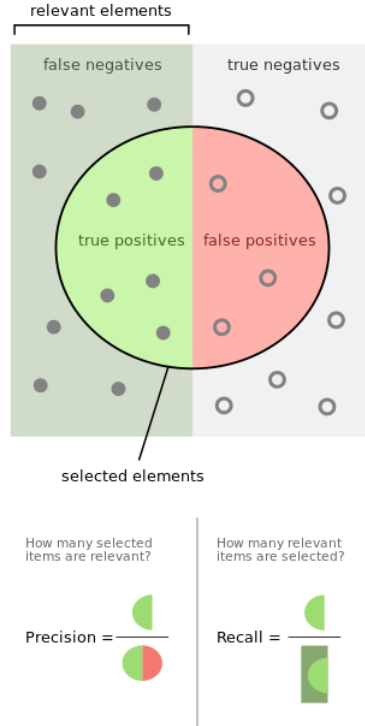


Figura 1: Retirado da Wikipedia.

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 simplesmente $\mathbb{P}(B|A)$. 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.
 - $\text{Precisão} = \text{Revocação} \times \frac{\mathbb{P}(\mathbf{X} \in 1)}{\mathbb{P}(D(\mathbf{x}) = 2)}$.
 - Existe um trade-off entre precisão e revocação: se aumentarmos uma métrica, a outra tem de diminuir.
3. Existem duas classes ou populações, 1 e 2, presentes nas proporções positivas π_1 e π_2 com $\pi_1 + \pi_2 = 1$. Suponha que $p_{i1} \approx 0$. O vetor aleatório contínuo $\mathbf{X} = (X_1, \dots, X_p)$ 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-classificação for também desbalanceado, com $c(2|1) \gg c(1|2)$, a estratégia anterior pode ser muito ruim. Obtenha o custo esperado de má-classificação (ECM) da regra anterior.