Cap 9: Exercícios 3, 4, 5, 6, 8

Cap 10: Exercícios 1, 2

Capıtulo 9

Classificacao

4. 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.

   - Precisão: $\mathbb{P}(\text{foto } \in \text{ gatos } | \text{ classificado como gato }) = \mathbb{P}(\mathbf{X} \in 1 | D(\mathbf{x}) = 1)$
   - 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 9.1, 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.

•Solução: V

•Solução: V

•Solução: F

•Solução: V

•Solução: F

5. Existem duas classes ou populacoes, 1 e 2, presentes nas proporcoes positivas π1 e π2 com π1+π2 = 1. Suponha que π1 ≈ 0. O vetor aleatorio contınuo X = (X1, . . . , Xp) com p variaveis

possui as densidades f1(x) e f2(x) quando o indivıduo pertence a populacao 1 ou 2, respectivamente. Seja

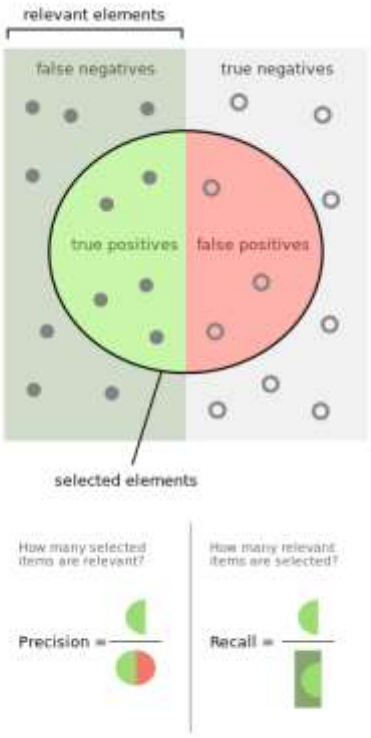

Figura 9.1: Retirado da Wikipedia.

c(1|2) o custo do erro de classificar erradamente no grupo 1 um indivıduo que seja do grupo 2. Analogamente, defina o custo do outro erro c(2|1). A regra de classificacao e representada pela funcao binaria D(X) que assume os valores 1 ou 2 dependendo do vetor aleatorio X cair ou nao na regiao R1 de classificacao no grupo 1. • Uma regra de decis˜ao que vai errar pouco sera atribuir a classe 2 a todo e qualquer item: D(X) ≡ 2 para todo valor de X. Obtenha a probabilidade de classificacao errada. A probabilidade e proxima de zero? • Se o custo de ma-clasificacao for tambem desbalanceado, com c(2|1) >> c(1|2), a estrategia anterior pode ser muito ruim. Obtenha o custo esperado de ma-classificacao (ECM) da regra anterior.

Solução: π1 e ECM = c(2|1)π1

Capıtulo 10

Teoremas Limite: LGN e TCL

1. Voce quer selecionar uma amostra para estimar a porcentagem θ de pessoas que vai votar num candidato X. Imagine que a resposta e uma v.a. X de Bernoulli com valores 1 e 0 (vai e nao vai votar, respectivamente) e a probabilidade de sucesso e θ. As respostas de n indivıduos serao X1, X2, . . . , Xn e voce vai estimar θ usando ˆθ = (X1 + . . . + Xn)/n, a proporcao amostral. Se voce asumir que as respostas sao variaveis aleat´orias i.i.d., determine o tamanho n da amostra necessario para que o erro de estimacao |ˆθ −θ| seja menor que 0.02 com probabilidade 0.99. Para isto, assuma que voce sabe que seu candidato esta estacionado entre 15% e 35% dos eleitores (baseado em outras pesquisas mais antigas). Esta e uma faixa de variacao enorme, muito pouco precisa, mas que voce esta bem

antigas). Esta é uma faixa de variação enorme, muito pouco precisa, mas que você está bem seguro de que ela contem a verdadeira proporcao de eleitores que votam no candidato em questao.

Solucao: Queremos encontrar n de forma que a probabilidade de ocorrer o evento |ˆθ − θ| < 0.02 seja 0.99. Isto ´e, queremos n de forma que P(|ˆθ−θ| < 0.02) < 0.99. Veja que ˆθ = (X1 +. . .+Xn)/n e portanto podemos usar o TCL. Temos Xi ∼ Bernoulli(θ) (bin´aria) independentes, com E(Xi) = θ e V(Xi) = θ(1 − θ). Assim, pelo TCL,

$$\begin{aligned}
\mathbb{P}(|\hat{\theta}-\theta|<0.02) &= \mathbb{P}(|\bar{X}-\theta|<0.02)\\
&= \mathbb{P}(-0.02<\bar{X}-\theta<0.02)\\
&= \mathbb{P}\left(-\sqrt{n}\frac{0.02}{\sqrt{\theta(1-\theta)}}<\sqrt{n}\frac{\bar{X}-\theta}{\sqrt{\theta(1-\theta)}}<\sqrt{n}\frac{0.02}{\sqrt{\theta(1-\theta)}}\right)\\
&\approx \mathbb{P}\left(-\sqrt{n}\frac{0.02}{\sqrt{\theta(1-\theta)}}<N(0,1)<\sqrt{n}\frac{0.02}{\sqrt{\theta(1-\theta)}}\right)
\end{aligned}$$

Sabemos que, no caso de uma v.a. $N(0,1)$, o valor a tal que $\mathbb{P}(-a < N(0,1) < a) = 0.99$ é igual $a = 2.58$ (pois, em R, o comando `qnorm(0.01/2)` retorna -2.575829). Assim, devemos ter $0.02\sqrt{n}/\sqrt{\theta(1-\theta)} = 2.58$. O valor de θ é desconhecido mas sabemos que ele está no intervalo $(0.15, 0.35)$. Como $\theta(1-\theta)$ é crescente com θ nesta região (cheque isto fazendo o gráfico desta função parabólica no intervalo $(0,1)$), tomamos o pior caso, em que $\theta = 0.35$, para calcular n. Queremos $0.02\sqrt{n}/\sqrt{0.35(1-0.35)} = 2.58$, o que implica em $n = 3785.827$. Basta tomar então uma amostra de tamanho 3786 para garantir o resultado.

2. No problema acima, usando uma amostra de tamanho n = 500, determine um intervalo da forma I = (ˆθ − c, ˆθ + c) tal que a probabilidade P(ˆθ − c ≤ θ ≤ ˆθ + c) seja aproximadamente igual ou maior que 0.95. Este tipo de intervalo e chamado de intervalo de confianca.

Solução: Um ponto fundamental é perceber que

$$\hat{\theta}-c\leq\theta\leq\hat{\theta}+c \iff -c\leq\hat{\theta}-\theta\leq c \iff |\hat{\theta}-\theta|\leq c$$

Assim, com $\hat{\theta} = (X_1 + \ldots + X_{500})/500$, queremos encontrar c tal que

$$\begin{aligned}
0.95 &= \mathbb{P}(-c\leq\hat{\theta}-\theta\leq c)\\
&= \mathbb{P}\left(-\sqrt{500}\frac{c}{\sqrt{\theta(1-\theta)}}\leq\sqrt{500}\frac{\hat{\theta}-\theta}{\sqrt{\theta(1-\theta)}}\leq\sqrt{500}\frac{c}{\sqrt{\theta(1-\theta)}}\right)\\
&\approx \mathbb{P}\left(-\sqrt{500}\frac{c}{\sqrt{\theta(1-\theta)}}\leq N(0,1)\leq\sqrt{500}\frac{c}{\sqrt{\theta(1-\theta)}}\right)
\end{aligned}$$

Mas, no caso de uma $N(0,1)$, temos $\mathbb{P}(-1.96 \leq N(0,1) \leq 1.96) = 0.95$ (verifique digitando `qnorm(0.05/2)`). Assim, devemos fazer $c\sqrt{500}/\sqrt{\theta(1-\theta)} = 1.96$. Como θ é desconhecido (mas dentro do intervalo $(0.15, 0.35)$), pegamos o pior caso ($\theta = 0.35$) para obter $c\sqrt{500}/\sqrt{0.35 \times 0.65} = 1.96$ o que implica em $c = 0.0418$.

In [ ]: