Questão 1

Relacione os conceitos às suas respectivas definições.

1. Medida da variabilidade de um estimador.
2. Valor do estimador calculado com os dados de uma amostra.
3. Distribuição de probabilidades de um estimador.
4. Quantidade, normalmente desconhecida, que especifica uma distribuição de probabilidades na população.
5. Função da amostra que representa valores plausíveis para o parâmetro desconhecido de interesse.

Alternativas:

(a) Estimativa

(b) Erro padrão

(c) Parâmetro

(d) Distribuição amostral

(e) Estimador

Solução: **1.(b); 2.(a); 3.(d); 4.(c); 5.(e).**

Questão 2

É muito comum a entropia cruzada, $H(p, q)$, ser adotada como função de perda em problemas de aprendizado de máquina. Em particular em problemas de classificação com duas classes (0 ou 1), usa-se a entropia cruzada binária. Veja um trecho da definição dessa entropia encontrada na página da *Wikipedia* (`https://pt.wikipedia.org/wiki/Entropia_cruzada`) abaixo.

> A entropia cruzada pode ser usada para definir uma função de perda no aprendizado de máquina e otimização. A verdadeira probabilidade p_i é o rótulo verdadeiro e a distribuição fornecida q_i é o valor previsto do modelo atual.

Tendo criado nossa notação, $p \in \{y, 1-y\}$ $q \in \{\hat{y}, 1-\hat{y}\}$, podemos usar entropia cruzada para obter uma medida de dissimilaridade entre p e q:

$$H(p,q) = -\sum_i p_i \log q_i = -y \log \hat{y} - (1-y)\log(1-\hat{y}).$$

Repare que essa função de perda, que se quer minimizar em problemas de aprendizagem de máquina, é (-1) vezes o ln de uma função de verossimilhança, que deve ser maximizada para a obtenção de estimadores de verossimilhança. A função de verossimilhança relacionada à essa entropia binária corresponde a que modelo probabilístico?

Alternativas:

(a) nenhuma das aprendidas em aula

(b) Bernoulli

(c) Binomial

(d) Poisson

(e) Normal

Solução: **Alternativa b.**

Tem-se que $H(p,q) = -y \log \hat{y} - (1-y)\log(1-\hat{y})$ é $(-1) * ln(L(p,q))$.
Assim, multiplicando-se por (-1) e aplicando a função inversa (e), obtém-se $L(p,q) = \hat{y}^y * (1-\hat{y})^{1-y}$.

Questão 3

Uma estratégia comum para solucionar o problema do aprendizado com conjunto de dados com classes desbalanceadas resume-se a métodos que visam balancear a distribuição das classes. Japkowicz (2000)[1]compara algumas abordagens para lidar com conjuntos com classes desbalanceadas e conclui que *under* e *over-sampling* são métodos efetivos para aprender nessas circunstâncias. *Under-sampling* resume-se a selecionar uma amostra aleatória da classe majoritária de modo a balancear ambas as classes. *Over-sampling* consiste em multiplicar algumas unidades da classe minoritária, com o mesmo intuito de balanceamento.

Suponha uma variável aleatória X com distribuição de Bernoulli, com parâmetro p. Seja o estimador da probabilidade de sucesso, $\hat{p}_{2n} = \sum_{i=1}^{2n} \frac{X_i}{2n}$, baseado numa amostra aleatória de tamanho $2n$. Analogamente, $\hat{p}_n = \sum_{i=1}^{n} \frac{X_i}{n}$ é o estimador da mesma probabilidade de sucesso, p, baseado numa amostra aleatória de tamanho n. É **incorreto** afirmar que:

Alternativas:

(a) A variância de $\hat{p}_n$ é maior do que a variância de $\hat{p}_{2n}$.

(b) Os erros quadráticos médios de ambos os estimadores são iguais às respectivas variâncias.

(c) O intervalo de confiança para p, com mesmo nível de confiança $1-\alpha$, terá a mesma amplitude para qualquer um dos dois estimadores que usarmos.

(d) A distribuição de ambos os estimadores converge para a distribuição Normal.

(e) Ambos os estimadores $\hat{p}_n$ e $\hat{p}_{2n}$ são não viciados para p.

Japkowicz, Nathalie. (2000). Learning from Imbalanced Data Sets: A Comparison of Various Strategies. In AAAI Workshop on Learning for Imbalanced Datasets, Menlo Park, CA. AAAI Press.

Solução: **Alternativa c.**

Note que:

$$\mathbb{E}(\hat{p}_n) = \mathbb{E}\left(\sum_{i=1}^{n} \frac{X_i}{n}\right) = \frac{1}{n}\left(\sum_{i=1}^{n} \mathbb{E}(X_i)\right) = \frac{1}{n}(n * p) = p;$$

$$\mathbb{E}(\hat{p}_{2n}) = \mathbb{E}\left(\sum_{i=1}^{2n} \frac{X_i}{2n}\right) = \frac{1}{2n}\left(\sum_{i=1}^{2n} \mathbb{E}(X_i)\right) = \frac{1}{2n}(2n * p) = p;$$

$$\mathbb{V}(\hat{p}_n) = \mathbb{V}\left(\sum_{i=1}^{n} \frac{X_i}{n}\right) = \frac{1}{n^2}\left(\sum_{i=1}^{n} \mathbb{V}(X_i)\right) = \frac{1}{n^2}\{n * [p * (1-p)]\} = \frac{p * (1-p)}{n};$$

$$\mathbb{V}(\hat{p}_{2n}) = \mathbb{V}\left(\sum_{i=1}^{2n} \frac{X_i}{2n}\right) = \frac{1}{4n^2}\left(\sum_{i=1}^{2n} \mathbb{V}(X_i)\right) = \frac{1}{4n^2}\{2n * [p * (1-p)]\} = \frac{p * (1-p)}{2n}.$$

Assim,

$$\mathcal{B}(\hat{p}_n) = \mathbb{E}(\hat{p}_{2n}) - p = 0;$$
$$\mathcal{B}(\hat{p}_{2n}) = \mathbb{E}(\hat{p}_{2n}) - p = 0;$$

$$EQM(\hat{p}_n) = \mathbb{V}(\hat{p}_n) + \mathcal{B}(\hat{p}_n) = \mathbb{V}(\hat{p}_n);$$
$$EQM(\hat{p}_{2n}) = \mathbb{V}(\hat{p}_{2n}) + \mathcal{B}(\hat{p}_{2n}) = \mathbb{V}(\hat{p}_{2n}).$$

Tem-se também (para n suficientemente grande):

$$\frac{\hat{p}_n - p}{\sqrt{\hat{p}_n(1-\hat{p}_n)/n}} \approx N(0,1);$$

$$\frac{\hat{p}_{2n} - p}{\sqrt{\hat{p}_{2n}(1-\hat{p}_{2n})/2n}} \approx N(0,1);$$

e os intervalos de confiança podem ser construídos usando-se as quantidades pivotais:

$$\hat{p}_n \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}};$$

$$\hat{p}_{2n} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}_{2n}(1-\hat{p}_{2n})}{2n}};$$

pode-se usar o intervalo de confiança conservador, e então:

$$IC_{\hat{p}_n}(1-\alpha) = \left(\hat{p}_n - z_{1-\alpha/2}\sqrt{\frac{1/4}{n}}; \hat{p}_n + z_{1-\alpha/2}\sqrt{\frac{1/4}{n}}\right);$$

$$IC_{\hat{p}_{2n}}(1-\alpha) = \left(\hat{p}_{2n} - z_{1-\alpha/2}\sqrt{\frac{1/4}{2n}}; \hat{p}_{2n} + z_{1-\alpha/2}\sqrt{\frac{1/4}{2n}}\right).$$

Note também que aumentar o valor de n diminui a amplitude do IC.

Questão 4

O número de barcos que chegam por dia em um porto secundário no estado do Rio de Janeiro (variável X) tem distribuição de Poisson de parâmetro λ. Numa amostra aleatória de tamanho 4, o total de barcos que chegaram nos 4 dias é igual a 20. Qual é a afirmação **incorreta**?

Alternativas:

(a) Os estimadores de máxima verossimilhança da média e da variância de X são iguais.

(b) O valor do estimador de máxima verossimilhança do número médio de barcos que chegam por dia no porto é igual a 5.

(c) Não é possível calcular o valor do estimador de máxima verossimilhança da média de X porque os valores individuais da amostra não foram fornecidos.

(d) Usar o Teorema Central do Limite (TCL) neste caso para obter o intervalo de confiança para λ é inadequado.

(e) O estimador de máxima verossimilhança de λ é não viciado.

Solução: **Alternativa c.**

Tem-se que:

$$X \sim Poisson(\lambda) : \text{número de barcos que chegam por dia no porto;}$$

$$\mathbb{E}(X) = \lambda = \mathbb{V}(X);$$

$$n = 4 \quad \text{e} \quad \sum_{i=1}^{4} X_i = 20.$$

As funções de verossimilhança e log-verossimilhança são, respectivamente:

$$L(\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda} * \lambda^{x_i}}{x_i!} \quad \text{e} \quad l(\lambda) = -n * \lambda + \sum_{i=1}^{n} xi * ln(\lambda) - ln(x_i!).$$

Assim,

$$\frac{\partial ln(\lambda)}{\partial \lambda} = -n * + \frac{\sum_{i=1}^{n} xi}{\lambda} = 0 \Rightarrow \hat{\lambda} = \frac{\sum_{i=1}^{n} xi}{n}.$$

Calculando $\mathbb{E}(\hat{\lambda})$,

$$\mathbb{E}\left(\frac{\sum_{i=1}^{n} xi}{n}\right) = \frac{1}{n}\left(\sum_{i=1}^{n} \mathbb{E}(X_i)\right) = \frac{1}{n}(n * \lambda) = \lambda.$$

Tem-se também, para n suficientemente grande:

$$\frac{\hat{\lambda} - \lambda}{\sqrt{\hat{\lambda}/n}} \approx N(0,1);$$

e o intervalo de confiança pode ser construídos usando-se as quantidades pivotais:

$$\hat{\lambda} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{\lambda}}{n}}.$$