Formulário [MAS]

1. Introdução

Medidas de Diversidade d

Λ <u>Menos diversidade</u> associa-se a um <u>menor risco de previsão</u> (classificação ou regressão) pelo que durante a aprendizagem se procura explicar/reduzir essa diversidade

Para uma variável **nominal**:

- A $d = 0$ quando todos os elementos de um conjunto de observações estão numa categoria
- A diversidade é $\boldsymbol{máxima}$ quando a distribuição é uniforme por categorias

Para uma variável **métrica**:

- A $d = 0$ quando todos os elementos de um conjunto de observações são iguais
- A diversidade aumenta quando as observações tendem a ser diferentes

Validação Cruzada

O recurso à validação cruzada permite obter estimativas mais realistas dos erros de aprendizagem.

Método V-Fold: a amostra original é particionada em V subamostras de dimensões iguais; cada submostra é excluída da aprendizagem à vez e os seus erros calculados com base no modelo estimado no treino:

$E_1^{(V)} \ldots E_{n_V}^{(V)}$ $\quad E_{n_V}^{(V)}$ estimativa de erro para cada observação n quando esta está na amostra de validação v (e <u>não está</u> na amostra de treino).

Regressão Alvo Métrico

Medidas de Diversidade

› Variância: $Var = S^2 = \frac{\sum_{n=1}^{N}(y_i-\bar{y})^2}{n}$

› Deviance: $DEV^{met} = \sum_{n=1}^{N}(y_i-\bar{y})^2 = Var \times (n-1)$

Residual Sum of Squares (RSS) or **Sum of Squares error (SSE)**

$$\sum(y_i - \hat{y}_i)^2$$

- Comparar ≠ resultados sobre o mesmo conjunto de dados
- Sensível à presença de *outliers*

Root Mean Squared Error (RMSE) $\sqrt{\frac{\sum(y_i-\hat{y}_i)^2}{n}}$

- Na mesma escala que y_i
- Ranking de modelos semelhante ao de medidas anteriores

Mean Absolute Error (MAE) $\frac{\sum|y_i-\hat{y}_i|}{n}$

- Comparar ≠ resultados sobre o mesmo conjunto de dados
- Não tão sensível à presença de *outliers* como medidas anteriores (baseadas nos quadrados dos erros)

Mean Absolute Percentage Error (MAPE) $\frac{\sum\left|\frac{y_i-\hat{y}_i}{y_i}\right|}{n} \times 100$

- O erro é comparado com a observação a que se refere

Ex: Um erro de 10 referido a uma observação cujo valor é 90 é diferenciado do mesmo erro referido a uma observação cujo valor é 900

Mean Squared Percentage Error (MSPE) $\frac{\sum\left(\frac{y_i-\hat{y}_i}{y_i}\right)^2}{n} \times 100\%$

- O erro é comparado com a observação a que se refere

(pseudo) R-squared | R^2

$$1 - \frac{\sum(y_i-\hat{y}_i)^2}{\sum(y_i-\bar{y})^2}$$

- Compara a diversidade residual com a diversidade inicial do alvo
- Compara o ajustamento obtido pelo modelo com o ajustamento proporcionado pela média
- $X\%$ → O modelo consegue explicar … % da variabilidade do valor médio de X
- Quando R^2 é negativo, significa que o modelo é pior a prever do que a média simples.

Nota: na Regressão Linear o mínimo é zero.

Relative Squared Error (RSE) $\frac{\sum(y_i-\hat{y}_i)^2}{\sum(y_i-\bar{y})^2}$

Relative Absolute Error (RAE) $\frac{\sum|y_i-\hat{y}_i|}{\sum|y_i-\bar{y}|}$

- Fornece o erro absoluto das previsões relativamente a um ***Modelo Naïve*** que usa a média como previsão para qualquer observação.

Classificação Alvo Nominal

Legenda

K - nº de classes / categorias da variável nominal;
n_k - nº de observações da classe / categoria k da variável
n - nº total de observações da variável nominal

Medidas de Diversidade

› Coeficiente de Gini: $G = 1 - \sum_{k=1}^{K}\left(\frac{n_k}{n}\right)^2$

CdG – Medida de Desigualdade de uma distribuição ($0 \le G \le \frac{1}{K}$

- Se $G = 0$ então corresponde à completa igualdade
- Se $G = 1/K$ então corresponde à completa desigualdade

› Deviance: $DEV^{nom} = -2 \times \sum_{k=1}^{K} n_k \log\left(\frac{n_k}{n}\right) = 2nH$

› Entropia*:

$$H = -\sum_{k=1}^{K}\frac{n_k}{n}\log\left(\frac{n_k}{n}\right)$$

› Entropia Normalizada

$$\text{Hn} = \frac{-\sum_{k=1}^{K}\frac{n_k}{n}\log\left(\frac{n_k}{n}\right)}{\log(K)}$$

* Ao considerar um evento impossível, convenciona-se que $0 * log(0) = 0$

Entropia (Medida de Dispersão para Dados Qualitativos)

- A Entropia de Shanon (H) permite avaliar em que medida a moda de uma variável qualitativa representa melhor ou pior os dados

Maior H, maior será o nº de estados possíveis, bem como a sua aleatoriedade.

- A Entropia (H) é a média da quantidade de informação que se associa a uma distribuição de frequências

A "quantidade de informação" associada a um acontecimento mais provável contém menos informação do que o de uma mais improvável

- Pode ser representado em $bits - 2$, $nats - e$ ou $decits - 10$.

› Hn varia entre 0 (**não há dispersão** - todas as obs. pertencem a 1 só classe/categoria (i.e., não há impureza - potencia boas previsão)

1 (**dispersão $máx$** - há impureza, o que dificulta a obtenção de boas previsões)

Matriz de Classificação *Confusion Matrix*

		Predicted	
			not
Actual		TP	FN
	not	FP	TN

Precision $\frac{TP}{TP+FP}$

- **Considerando o total das observações que se preveem como 1 (/positivos)** indica a proporção de observações corretamente classificadas.

Recall/ Sensitivity $\frac{TP}{TP+FN}$

- **Considerando o total de observações 1 (/positivos)** indica a proporção de corretamente classificadas.

Specificity $\frac{TN}{TP+TN}$

- **Considerando o total de observações 0 (/negativos)** indica a proporção corretamente classificadas.

Elevada Especificidade → *Menos Provável a ocorrência de FP*

F1 - Score $2 \times \frac{Recall \times Precision}{Recall + Precision} = \frac{TP}{TP+\frac{1}{2}(FP+FN)}$

Accuracy

$$\frac{TP+TN}{TP+FN\ FP+TN)} = \frac{\sum diagonal}{n}$$

- % de obs. corretamente classificadas

Índice de Huberty $\frac{\sum_{k=1}^{K}\frac{n_{kk}}{n}-p^{\text{def}}}{1-p^{\text{def}}}$

- O modelo preditivo desenvolvido permite melhorar a previsão em $X\%$ em relação ao modelo preditivo por defeito (dado pela classe modal)

$\sum_{k=1}^{K}\frac{n_{kk}}{n}$ **% de corretamente classificados** (= Accuracy)

p^{def} **% de classe maioritária** (sum. da linha com mais obs / n)

Classificador Naïve Bayes

- O classificador Naïve Bayes tem como base o *Teorema de Bayes* usando como pressuposto a <u>independência dos preditores</u>:

$$P(c_k \mid x) = \frac{P(x \mid c_k)P(c_k)}{P(x)} = \frac{P(c_k)\prod_{j=1}^{J}P(x_j \mid c_k)}{P(x)} \propto P(c_k)\prod_{j=1}^{J}P(x_j \mid c_k)$$

é proporcional a

- O classificador **Naïve Bayes** pode ser usado como referência para comparar o desempenho de outros métodos.

Preditores Qualitativos

$$P(c_k \mid x) \propto P(c_k) \cdot \prod_{j=1}^{J}\prod_{l \in \{j_1,\ldots,j_{Lj}\}} P(x_j = l \mid c_k)$$

Onde $\prod_{j=1}^{J}P(x_j = j_1 \mid c_k)$ é aproximado por produto de frequências relativas associadas a categorias de preditores intra-classe alvo c_k

```
Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = x, y = y)

A-priori probabilities:
y
     play don't Play
 0.6428571  0.3571429

Conditional probabilities:
            OUTLOOK
y              overcast      rain     sunny
   play       0.4444444 0.3333333 0.2222222
  don't Play 0.0000000 0.4000000 0.6000000
```

P(sunny | play)

Preditores Contínuos

$$P(c_k \mid \underline{x}) \propto P(c_k)\prod_{j=1}^{J}P(x_j \mid c_k)$$

Onde $\prod_{j=1}^{J}P(x_j \mid c_k)$ é aproximado por produto de Gaussianas:

$$\prod_{i=1}^{J}\frac{1}{\sqrt{2\pi S_k^2}}\exp\left\{-\frac{(x_j-\bar{x}^k)^2}{2S_k^2}\right\}$$

```
A-priori probabilities:
y
    setosa versicolor  virginica
 0.3333333  0.3333333  0.3333333

Conditional probabilities:
            Sepal.Width
y              [,1]      [,2]
  setosa      3.428 0.3790644
  versicolor  2.770 0.3137983
  virginica   2.974 0.3224966
```

Mean in col. 1
Std. Dev in col. 2

Função Densidade de Probabilidade da Normal

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-(x-\mu)^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\}$$

$-\infty < x < \infty$ $\quad E[X] = \mu \to \bar{x}$ $\quad Var(X) = \sigma^2$

K-Nearest Neighbour (KNN)

KNN (Não Paramétrico) VS Abordagem Paramétrica

VANTAGENS	DESVANTAGENS
Pode ser usado em Regressão e Classificação	É necessário parametrizar o valor K
É capaz de lidar com funções alvo complexas	Custos elevados em termos computacionais: armazenamento e processamento
É bastante preciso e simples de usar	O KNN é muito sensível à presença de atributos irrelevantes
Não considera pressupostos sobre os dados	O KNN é naturalmente dependente da escala do conjunto de dados e das medidas de distância utilizadas
Pode ser usado para uma ampla variedade de problemas.	

KNN | Passo a Passo

1. Indicar K, o número de vizinhos, e a medida de distância

O parâmetro K irá determinar o modo de aprendizagem – para uma nova observação, ele define o nº de observações mais próximas que serão a base da previsão do valor do alvo.

No entanto a escolha de um valor K específico poderá ter um efeito drástico nos resultados obtidos.

Quando K tem um valor reduzido o método tende a ter baixo enviesamento, mas Var muito alta. Conforme K cresce, o método torna-se menos flexível, apresentando uma variância reduzida, mas enviesamento elevado. Naturalmente k não pode ser superior a $n-1$…

A experimentação com diversos valores de K num contexto treino-teste ou de validação cruzada pode apoiar a seleção de K.

2. Calcular a distância entre um exemplo e os exemplos do conjunto de dados

As implementações do KNN no *R* recorrem, usualmente, à medida de distância *Euclideana* (eventualmente a distâncias de *Minkowski* das quais a *Euclideana* é um caso particular com $p = 2$)

$$d(x,y) = \sum_{i=1}^{n}\{|x_i - y_i|^p\}^{1/p} \quad \to \quad d(a,b) = \sqrt{(a_x-b_x)^2 + (a_y-b_y)^2}$$

Nota: A integração de preditores qualitativos poderá ser feita mediante codificação *dummy* $(n-1)$.

- Para que os preditores tenham protagonismo similar na previsão será necessário **NORMALIZAR**

$min - max$ (+ comum – resulta em valores [0,1])

```
normalize <- function(x){
  return ((x -min(x)) / (max(x)-min(x)))}
```

$média - desvio\ padrão$

```
standardize <- function(x){
  return ((x -mean(x)) / sd(x))}
```

Nota: Quando normalizamos temos de fazer para todas as variáveis, mesmo as binárias!

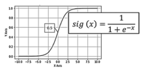
3. Ordenar as distâncias calculadas (ordem crescente)

4. Obter os valores do alvo para as K observações mais próximas do exemplo

Regressão Logística

Métodos de Classificação

- No modelo de regressão logística para <u>classificação binária</u> é usada a função **sigmóide** ou **logística**
  › *Logistic Regression (binary response)*



- No modelo de regressão logística, para <u>classificação em múltiplas classes</u>, usa-se a função **softmax**
  › *Multinomial Logistic Regression ou Softmax Regression*

O objetivo de aprendizagem é

<u>Minimizar</u> *ce* (*Entropia Cruzada*)

$$ce = H(y,\hat{y}) = -\sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}\log(\hat{y}_{ik})$$

ou, <u>Minimizar</u> a medida *Residual Deviance*

$$RESID\ DEV^{nom}(y_i,\hat{y}_i) = \sum_{i=1}^{n}2\left[\sum_{k=1}^{K}y_{ik}\,ln\left(\frac{y_{ik}}{\hat{y}_{ik}}\right)\right] = -2log(L) = 2 \cdot ce$$

ou, <u>Máximizar</u> o **logaritmo** da *Função de Verosimilhança*

$$L = \prod_{i=1}^{n}P(Y = y_i \mid X = \underline{x}_i) = \prod_{i=1}^{n}\pi(\underline{x}_i)^{y_i}[1-\pi(\underline{x}_i)]^{1-y_i}$$

- Quando o **ajustamento** do modelo for *perfeito* estes **indicadores** serão **nulos**

Função Sigmóide ou Logística

- Sendo $\boldsymbol{Y}$ **uma var. dependente binária** e $\boldsymbol{X}$ um preditor considerese

$$P(x \in C_1 \mid x) = P(Y = 1 \mid x) = \pi(x)$$
$$P(x \in C_2 \mid x) = P(Y = 0 \mid x) = 1 - \pi(x)$$

O Modelo de Regressão Logística ou *Modelo Logit* resulta de uma relação não linear entre a probabilidade $\pi(x)$ e x, modelada pela função **sigmóide**:

$$\hat{\pi}(x) = sig(x) = \frac{1}{1+e^{-[\beta_0+\beta x]}} = \frac{e^{\beta_0+\beta x}}{1+e^{\beta_0+\beta x}}$$

Nota: a combinação linear $[\beta_0 + \beta x]$ pode ser alargada a + preditores.

Função Logit

A função *logit* é a **inversa da função logística**:

$$logit(\hat{\pi}(x)) = ln\left[\frac{\hat{\pi}(x)}{1-\hat{\pi}(x)}\right] = ln\left(e^{\beta_0+\beta x}\right) = \beta_0 + \beta x$$

- *log-odds* or *logit* considerando C_2 como *Classe de Referência*
- *logit* de $\pi(x)$ pode ser qualquer número real (embora $\pi(x) \in [0,1]$)

Odds Ratio → <u>Razão de chances</u> (Comparando probabilidade de sucesso e probabilidade de fracasso)

	odd
x	$\frac{P(play \mid HUMIDITY = x)}{P(don't\ play \mid HUMIDITY = x)} = exp(7.212) * exp(-0.08x)$
$x+1$	$\frac{P(play \mid HUMIDITY = x+1)}{P(don't\ play \mid HUMIDITY = x+1)} = exp(7.212) * exp(-0.08x) * exp(-0.08)$

$\frac{\text{odds depois de um incremento unitário do preditor}}{\text{odds antes de um incremento unitário do preditor}} = exp(-0.08)$

$exp(-0.08) = 0.923 < 1$
Há uma diminuição na "razão de chances" de jogar que deriva do aumento unitário da humidade

Escolha de Múltiplos Preditores

- É importante ter em conta que a omissão de preditores importantes (correlacionados) pode "confundir" a interpretação
- A questão da *Multicolinearidade* volta a colocar-se, tal como na regressão linear múltipla
- Notar que um preditor qualitativo deverá ser codificado com o auxílio de *variáveis dummy*.

< 0.4 – Baixa $\quad 0.4 \le Corr \le 0.69$ –Moderada $\quad \ge 0.7$ –Alta

Medidas de Multicolinearidade

Tolerance (TOL) para X_j: $1 - R_j^2$ em que R_j^2 é o o coeficiente de determinação relativo à RLM de X_j sobre os restantes preditores; assim, quanto maior a TOL (menor o R_j^2) melhor;

→ É geralmente aceite que existe uma forte multicolinearidade se $TOL_j < 0.1$.

Variance Inflation Factor (VIF): o inverso de *TOL*, que também quantifica quanto a variação do estimador do coeficiente é inflacionada pela presença de multicolinearidade; assim, quanto menor o *VIF*, melhor;

→ Assume valores ≥ 1 e é geralmente aceito que há uma forte multicolinearidade se $VIF\hat{\alpha}_j > 10$.

Medidas de Avaliação da Regressão

- Accuracy
- Índice de Huberty $\frac{\text{accuracy} - p^{\text{def}}}{1-p^{\text{def}}}$

Interpretação: A capacidade preditiva desta *regressão* traduz uma melhoria relativa de $X\%$ face ao diferencial entre a capacidade preditiva de 100% e a capacidade preditiva de modelo que propõe classificar todas as observações na classe modal.

- Regressão Linear $\to \boldsymbol{R^2} = 1 - \frac{\sum y_i - \hat{y}_i}{\sum y_i - \bar{y}_i} = 1 - \frac{SSE}{Deviance}$

- R^2 *de Hosmer and Lemeshow* → $R_{HL}^2 = \frac{(-2*\log(L^0))-(-2*\log(L))}{-2*\log(L^0)}$ sendo L^0 a verosimilhança sem preditores;

A medida R_{HL}^2 mede uma melhoria relativa na *Residual Deviance*

Pretende-se então um valor máximo para R_{HL}^2 que indica uma capacidade preditiva máxima para o modelo considerado

A diferença $(-2 * \log(L^0)) - (-2 * \log(L))$:

- $-2 * log(L^0)$ é a *Residual Deviance* do modelo sem preditores (apenas com constante)
- $-2 * log(L)$ é a *Residual Deviance* do modelo com preditor(es)

- Critério *AIC (Akaike's criterion)* → $AIC = -2log(L) + 2\theta$ em que θ é o nº de parâmetros a estimar (nº de valores no *output*)

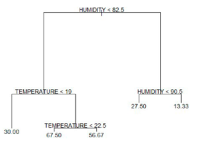
É um critério da *Teoria de Informação* que faz um *trade-off* entre o ajustamento pela *Função de Verosimilhança* e a *Complexidade do Modelo*:

Pretende-se então um **valor mínimo para *AIC***

Algoritmo CART

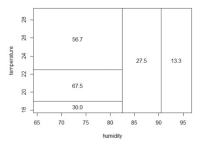
CART - Classification and Regression Trees

A construção das **Árvores de Decisão** é baseada num procedimento recursivo que divide os dados para conquistar (*greedy*) uma boa previsão - são modelos *explicativos* e *preditivos*.



No **nó raiz** estão todas as observações.

Nos **nós folha** são feitas as previsões.



- Lidam com variáveis explicativas quantitativas e qualitativas, *não sendo necessária normalização*
- Lidam com *Valores Omissos* e *Multicolinearidade*

- **Ponto de partida:** o conjunto de treino é colocado no nó raiz da árvore.
- **Processo:** o nó raiz é dividido primeiro em 2 nós descendentes, dando origem a um novo nível da árvore (isso é repetido nos nós descendentes)
- **Objetivo:** Reduzir a diversidade da var. alvo

No lim queremos $S^2 = 0$
(Todas as possibilidades são =)

CART | Construção

1. Um critério para decidir qual é a melhor **ramificação** de um nó
2. Um método de **previsão** de cada nó da folha
3. Regras para decidir **parar o processo de ramificação** num nó folha
4. Medidas apropriadas de **desempenho do modelo**
5. Critérios para **cortes de ramos** da árvore (poda)

No algoritmo **CART**, as **ramificações são binárias**:

- Se a variável explicativa for <u>métrica</u> ou tiver K **valores ordenados**, $K-1$ **possíveis divisões** serão consideradas com base nos pontos médios entre esses valores ordenados.
- Se a variável explicativa <u>nominal</u> tiver K **categorias**, $2^{K-1}-1$ **divisões possíveis** serão consideradas

De entre todas as ramificações possíveis num nó, seleciona-se a que proporciona o **maior decréscimo de diversidade** – *Deviance* (**R** e **C**) *ou Coeficiente de Gini* (**C**)

Exemplo: `Play`

Árvore de **Regressão**

(com 2 preditores)

```
node), split, n, deviance, yval
* denotes terminal node
1) root 14 11580.0 36.79
  2) HUMIDITY < 82.5 7  3843.0 52.14
    4) TEMPERATURE < 19 2  1800.0 30.00 *
    5) TEMPERATURE > 19 5   670.0 61.00
      10) TEMPERATURE < 22.5 2    12.5 67.50 *
      11) TEMPERATURE > 22.5 3   516.7 56.67 *
```

R^2

Ramificação do nó 5):

→ O maior decréscimo de diversidade no nó 5) foi obtido com esta ramificação

DEV(O1)= ...=11580.0
...
DEV(O5)= (50-61)**2+(70-61)**2+ (75-61)**2+ (45-61)**2 + (65-61)**2=670
DEV(O10)= (70-67.5)**2 + +(65-67.5)**2=12.5
DEV(O11)= (50-56.7)**2 + +(75-56.7)**2+ (45-56.7)**2 =516.7

➢ O decréscimo de diversidade na ramificação de nó 5)
(5/14)*670-[(2/14)*12.5+ (3/14)*516.7]=126.7786

Árvore de **Classificação**

(com ajustamento perfeito)

```
node), split, n, deviance, yval, (yprob)
* denotes terminal node
1) root 14 18.250 Play ( 0.3571 0.6429 )
  2) OUTLOOK: rain, sunny 10 13.860 Play ( 0.5000 0.5000 )
    4) HUMIDITY < 82.5 5  5.004 Play ( 0.2000 0.8000 )
      8) TEMPERATURE < 19 1  0.000 Don't Play ( 1.0000 0.0000 ) *
      9) TEMPERATURE > 19 4  0.000 Play ( 0.0000 1.0000 ) *
    5) HUMIDITY > 82.5 5  5.004 Don't Play ( 0.8000 0.2000 )
     10) HUMIDITY < 95.5 4  0.000 Don't Play ( 1.0000 0.0000 ) *
     11) HUMIDITY > 95.5 1  0.000 Play ( 0.0000 1.0000 ) *
  3) OUTLOOK: cloudy 4  0.000 Play ( 0.0000 1.0000 ) *
```

Em tree o cálculo de Deviance recorre ao logaritmo natural

ln

Accuracy=1

P(Don't Play) P(Play)

9/14

➢ D(O1)= -2*(9*ln(0.643)+5*ln(0.357))=18.249
➢ D(O2)=-2*(5*ln(0.5)+5*ln(0.5))=13.863
➢ D(O3)=0
➢ D(O4)= D(O5)= -2*(5*0.2*ln(0.2)+5*0.8*ln(0.8))= 5.004

O decréscimo de diversidade:

➢ **Na ramificação de nó 1:**
(14/14)× 18.249-((10/14)*13.863+ (4/14)*0)=8.346857

➢ **Na ramificação do nó 2:**
(10/14)* 13.863 -2*(5/14)* 5.004=6.328

3. Regras de Paragem

- Um decréscimo $mín$ de diversidade da var. dependente
- Um n° $mín$ de observações em nós "pais" e nós "filhos"
- Um n° $máx$ de níveis na árvore

4. Medidas de Desempenho

Regressão

Deviance: $\sum_{O \in A_F} \sum_{i \in O} \left(y_i - \bar{y}^{(O)}\right)^2$

No final, a *Residual Deviance* será quantificada pela *Deviance* nos nós O pertencentes ao conjunto de nós folha (A_F)

Classificação

Accuracy:

- Num nó O: $P_c(O)$ - % de obs. corretamente classificadas em O
- Na árvore A (soma, ponderada pela frequência relativa dos nós, da accuracy em todos os nós folha/conjunto A_F):

$$P_c(A) = \sum_{O \in A_F} p(O) P_c(O)$$

5. Critérios para Cortes de Ramos da Árvore (Poda)

A poda segue um critério de **custo** (precisão) levando em consideração a **complexidade** (número de nós folha). – *<u>Precaver o overffiting</u>*

Ponto de Inflexão no gráfico $y = dev \mid x = size$

Exame 2021

1. a) `knn.reg(Boston_n,y=Boston_n$medv, k=1,algorithm="brute")`

- Validação *one-hold-out* → n é o n° de *folds*
- Como "brute" o KNN usa o vizinho mais próx. (e não a própria obs).

c) Calcule o ***SSE*** – Sum of Square Error - (ou ***RSS*** – Residual Sum of Square)

$$SSE = \sum (y_i - \hat{y}_i)^2 \qquad R^2 = \frac{\sum y_i - \hat{y}_i}{\sum y_i - \bar{y}_i} = 1 - \frac{SSE}{Deviance} = \dots$$

d) Normalização $0-1$ do valor x_i da variável X:

$$\frac{x_i - min(X)}{max(X) - min(X)}$$

Distância Euclidiana entre $x = (x_1, \dots, x_n)$ e $y = (y_1, \dots, y_n)$:

$$d(x,y) = \sqrt{(x_1-y_1)^2 + (x_2-y_2)^2 + \dots + (x_n-y_n)^2} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

- A pertinência de normalizar no **KNN** é a **amplitude** das variáveis.

2. a) $Hn = \dots\dots = 0.87$ → O valor está muito próximo de 1 indicando *dispersão elevada* pelas 3 categorias de `origin`.

c) O modelo de **RL Multinomial** considera equações baseadas na função **softmax** para determinar as **probabilidades** das classes alvo;

Para obter o AIC soma-se **2** vezes **n° parâmetros estimados** à *Residual Dev*.

NOTA: O `intersept` é a classe de referência - Não aparece no *output*

3. a) A **Árvore de Classificação** aprende sobre **todas as** observações do *dataset* (basta ver se o n na `root` – onde está a *Deviance* da **variável alvo** tmb – se é igual ao n do *dataset*)

c) Índice de Huberty $= \frac{\text{Accuracy} - \text{\% Classe Modal}}{1 - \text{\% Classe Modal}} = \dots \approx 0.497$

A capacidade preditiva desta árvore traduz uma melhoria relativa de **49.7%** face ao diferencial entre a capacidade preditiva de **100%** e a capacidade preditiva de modelo que propõe classificar todas as observações na classe modal.

Exame 2022

1. b) $DEV^{nom} = -2 \times \left[n_{\text{"No"}} \times log\left(\frac{n_{\text{"No"}}}{n}\right) + n_{\text{"Yes"}} \times log\left(\frac{n_{\text{"Yes"}}}{n}\right) \right]$

c) • No **KNN** ter atenção à multicolinearidade

• A constituição das amostras de treino/teste **não** recorreu a um processo **estratificado**.

d) Standardização μ/σ do valor x_i da variável X: $\frac{x_i - \bar{X}}{s_X}$

$$Z = \frac{144 - 103{,}53}{112{,}07} = 0{,}36 \rightarrow$$

Para a observação 1 do conjunto de teste, a *experiência prévia*, em meses, encontra-se acima da média (observada neste conjunto); o desvio face à média deste valor observado é menor que o desvio padrão.

c) <u>Decréscimo de Diversidade:</u>

```
Regression tree:
snip.tree(tree = rtree_large.employee, nodes = c(7L, 4L, 12L,
5L, 26L))
Number of terminal nodes:  6
Residual mean deviance:  72.57 = 20760 / 286
Distribution of residuals:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-22.420  -4.584  -1.399   0.000   3.266  48.620

> rtree.employee
node), split, n, deviance, yval
      * denotes terminal node

 1) root 292 74930.00 34.61
   2) educ < 15.5 221 12710.00 28.00
     4) gender: f 109  2761.00 24.52 *
     5) gender: m 112  7350.00 31.38 *
   3) educ > 15.5 71 22440.00 55.20
     6) educ < 17.5 46  9439.00 47.18
      12) gender: f 17  1574.00 38.19 *
      13) gender: m 29  5684.00 52.45
        26) minority: No 26  4396.00 54.64 *
        27) minority: Yes 3    87.18 33.50 *
     7) educ > 17.5 25  4588.00 69.96 *
```

Nó 3) deviance= 22440 (n° obs.= 71)

Nós descendentes de 3)

Nó 6) deviance= 9439 (n° obs.= 46)

Nó 7) deviance= 4588 (n° obs.= 25)

Decréscimo: $(71/292) * 22440 - ((46/292) * 9439 + (25/292) * 4588 = 3576.527$

d) • Para efetuar uma ramificação `rtree_large.employee` ~~exige o valor~~ 0.0001 como decréscimo mínimo da deviance

F – $mindev$ –> $0.0001 \times Deviance\ inicial$ **é o mínimo valor de decréscimo da *Deviance***

- A *Residual Deviance* que se associa ao modelo obtido é **20760** **V**
- As previsões em `rtree.employee` obtêm-se apenas em 6 nós **V – é só nos nós folha**
- É no nó ~~folha~~ 3) que se observa o maior decréscimo da deviance **F – 3) não é um nó folha**

2. b) Índice de Huberty $= 0$ → A capacidade preditiva do modelo nada acrescenta à que se obtém mediante a mera afetação à classe modal **"No"**.

c) Calcule a probabilidade de um indivíduo na categoria `"Manager"` e com `"Salary=40000"` estar na classe minoritária (`"Yes"` de `"minority"`) de acordo com o modelo **Naïve Bayes** estimado (e indique a classe em que será classificado de acordo com o mesmo modelo).

```
Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = x, y = y)

A-priori probabilities:
y
       No       Yes
0.7945205 0.2054795

Conditional probabilities:
     jobcat
y      Clerical Administrative    Manager
  No     0.741      0.034          0.224
  Yes    0.850      0.117          0.033

     salary
y         [,1]      [,2]
  No  36511.64 17012.457
  Yes 27261.67  8225.067
```

P(minority=**Yes** | jobcat=Manager e Salary=40000) ∝ **é proporcional a**

P(m=Yes) × P(j=Manager | m=Yes) × P(S=40000 | m=Yes)=

0.205 × 0.033 × 0.000022

já que $\phi(40000; \mu = 27261.67; \sigma = 8225.067) = \frac{1}{8225.067\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left[\frac{40000 - 27261.67}{8225.067}\right]^2\right\} = 0.0000222$

P(minority=**No** | jobcat=Manager e Salary=40000) ∝ **é proporcional a**

P(m=No) × P(j=Manager | m=No) × P(S=40000 | m=No)=

$0.795 \times 0.224 * (40000;\ \mu = 36511.64;\ o = 17012.457) = 0.795 * 0.224 * 0.000021$

já que $\phi(40000; \mu = 36511.64; \sigma = 17012.457) = \frac{1}{17012.457\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left[\frac{40000 - 36511.64}{17012.457}\right]^2\right\} = 0.000021$

Logo, P(minority=**Yes** | jobcat=Manager e Salary=40000) $= \frac{0.205*0.033*0.000022}{0.205*0.033*0.000022 + 0.795*0.224*0.000021} = \frac{0.0008}{0.0708} = 0.038$

P(minority=**No** | jobcat=Manager e Salary=40000) $= \frac{0.795*0.224*0.000021}{0.205*0.033*0.000022 + 0.795*0.224*0.000021} = \frac{0.07}{0.0008 + 0.07} = 0.962$

∴ Pelo que a obs. se classifica em **`minority = No`**

3. a) Escolha dos Preditores no **KNN**

• Ao procurar 2 preditores não correlacionados entre si tem-se em conta o pressuposto implícito no uso da *<u>Distância Euclidiana</u>* no **KNN**:

Pressuposto de dimensões não correlacionadas/ortogonais

b) Ter atenção à classe **`positive`** ⚠

	Yes	No		Yes	No
Yes	10	17	Yes	TP	FN
No	32	99	No	FP	TN

$$Sensitivity = \frac{TP}{TP+FN} = \frac{10}{10+17} = 0.370$$

A capacidade de prever corretamente casos que são positivos é **37%**.

$$Precision = \frac{TP}{TP+FP} = \frac{10}{10+32} = 0.238$$

De entre os casos que se preveem positivos registam-se **23.8%** de previsões corretas.

A ***sensitivity*** é maior que a ***precision***, i.e., o modelo favorece a previsão correta de casos efetivamente positivos face à capacidade de prever positivos com precisão

c) Os dados normalizados entre **0** e **1** e entre a *média* e o *desvio padrão*, resultam em vizinhos mais próximos **diferentes**;

d) Que valor da distância é considerado pelo **`knn.employee`**, entre a obs. 2 de teste e a primeira observação do conjunto de treino?

Considerando os preditores salary e jobtime (estandardizados) temos os dados seguintes:

<u>2ª obs. de Teste</u>: $salary = -0.742$ $jobtime = 1.621$

<u>1ª obs. de Treino</u>: $salary = -0.521$ $jobtime = -1.471$

Distância Euclidiana: $\sqrt[2]{(-0.742 - (-0.521))^2 + (1.621 - (-1.471))^2} = 3.100$

e)

```
knn(employee_train_s[,predictors],employee_test_s[,predictors],
          employee_train_s$minority, 3, prob = TRUE)
```

- A **distância Euclideana** neste **KNN** é feita com uma obs. de treino e uma de teste.
- **Não** foi feita ***one-hold-out*** *cross-validation* → Só se aplica quando não se põe o conjunto de teste

4. a) $R^2 = 1 - \frac{Deviance/RSS}{sd^2_{inicial} \times n}$

b)

```
rtree_large.employee <-tree(salary/1000 ~ gender+educ+minority, data=employee_train,
                      control=tree.control(nrow(employee_train),
                                           mincut = 2,
                                           minsize = 4,
                                           mindev = 0.0001), split = "deviance")
```

⚠ Ver como é que *target* está na tree ⚠

Para ver o valor estimado pela **Árvore de Regressão** temos de ver o **`yval`** e converter (se necessário

Exame 2023

1. b) • Cálculo da **Entropia** <u>não se considera NAs</u>

2. b) $Accuracy = (87 + 288)/(87 + 188 + 42 + 288) = 375/605 = 0.6198$

(Sim) Accuracy é maior que *Default accuracy* que resulta da afetação de todas as obs. à classe modal, cuja freq. relativa é 330/605=0.5454

c) Cálculos da probabilidade $P(ii2.3 = 4\ e\ ii2.11 = 3 \mid ii1\_0.1 = \text{"}Não\text{"})$ - **NB**.

(atendendo à independência dos preditores, que se pressupõe)

$$\dots = P(12.3 - 41\ 111\ 0.1 - Não) \times P(2.11 - 31\ 1\ 0.1 - Não) = **$$

Tendo em conta que as distribuições condicionais acima são modeladas por Gaussianas com fdp $f(x) = exp\{-()\}$ com os seguintes parâmetros referidos à classe <u>Não</u> (indicados no modelo NB)

ii2.3 $\mu = 3.221 \mid \sigma = 0.71$ **ii2.11** $\mu = 3.713 \mid \sigma = 0.548$

$$** = \frac{1}{0.71\sqrt{2 * 3.1416}} \exp\left\{-\frac{1}{2}\left(\frac{4 - 3.221}{0.71}\right)^2\right\} * \frac{1}{0.548\sqrt{2 * 3.1416}} \exp\left\{-\frac{1}{2}\left(\frac{3 - 3.713}{0.549}\right)^2\right\} =$$

$$= 0.3077851 * 0.3122714 = 0.09611248$$

3. a) • A diferença entre o *AUC* e *Residual Deviance* **não** corresponde ao n° de parâmetros estimados (corresponde a 2x esse número)

b) MRL estimado

$$log\left(\frac{P(sim, investe\ em\ FIF)}{1 - P(sim, investe\ em\ FIF)}\right) = -5.1025 + 0.5386ii2.3 + 0.9115ii2.11$$

Interpretação de 0.5386: Quando $ii2.3$ aumenta 1 unidade (sendo assim atribuído mais um grau de importância ao seguro de riscos de exploração quando se pondera a decisão de investir em FIF), há uma variação de $exp(0.5386) = 1.7136$ (i.e. um aumento) da razão de chances de investir em FIF face à de não investir

c) Probabilidade de a 1° Observação pertencer à classe **"Não"**

$$P(Não \mid ii2.3 = 4, ii2.11 = 3) = 1 - \frac{1}{exp\{-[-5.1025 + 0.5386 + 4 + 0.9115 * 3]\}} =$$
$$= 1 - 0.4468 = 0.5532$$

4. a) Em que **nó folha se classifica**

A 1° obs. do conjunto de teste tem ii1_0.1= $Não$, 2.3 = 4, 12.10 = 3, 12.11 = 3, 2.13 = 3, 2.14 = 1

i12.11 = 3 < 3.5 ⇒ Segue para o no 2)

i12.13 = 3 < 3.5 ⇒ Segue para nó 4)

ii2.3 = 4 > 2.5 ⇒ Segue para nó 9)

i12.14 = 1 < 1.5 ⇒ Segue para nó 18) (nó folha onde é classificada em <u>Sim</u>)

Originalmente il1_0.1= $Não$ pelo que esta obs. é incorretamente classificada.

c) Calcula-se o ***Indice de Huberty*** para o conjunto treino e teste

Face à melhoria possível sobre a ***Accuracy*** obtida mediante a classificação na classe modal, a **AC** proporciona uma melhoria relativa de **24.4%** na amostra de treino e de **7.46%** na amostra de teste; é uma diferença substancial que indicia **sobreajustamento.**

5. a) KNN - Com **Treino/Teste**

Para cada observação da amostra de **teste** forma considerados os vizinhos mais próximos na amostra de **treino**

b) "Os dados dos preditores deveriam ter sido ***normalizados*** antes de proceder à análise do **KNN**"

• Não seria necessário já que todos os preditores se encontram na mesma escala
(1-nada importante... 4-muito importante)

c) ***RSE*** *(Relative Squared Error)* correspondente às estimativas obtidas pelo **KNN** e interprete.

$$RSE = \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} = \frac{105988}{21.65^2 \times (226\text{-}1)} = \frac{105988}{105462.6} = 1.004982$$

A soma dos erros quadráticos associados ao modelo é ligeiramente superior à soma dos erros quadráticos em torno da média (RSE ligeiramente acima de 1) o que indica que o KNN tem um desempenho pior que o simples modelo que adota como previsão a média dos valores alvo.

ii1-Em igualdade de condições de risco e rentabilidade, com que probabilidade investiria num Fundo de Investimento Florestal?