Aprendizagem Automática
Sistemas de Classificação

$\mathbb{G}$. Marques

Motivação:

- A classificação é um processo de categorização ou identificação em que objectos, ideias, seres, etc, são agrupados por classe.
- Em aprendizagem automática a classificação é um processo **supervisionado**. O objectivo é associar a uma nova observação uma classe de um conjunto pré-definido de classes, usando para tal **um conjunto de treino** com exemplos em que se conhece a classe.
- Para poder usar algoritmos de classificação, os dados (objectos, ideias, seres, etc...) têm que ser representados por um conjunto finito de **características**. Cada característica corresponde a uma propriedade mensurável dos dados. Cada objecto é representado por um vector, em quem cada dimensão contém o valor de uma característica.
- Para avaliar o desempenho dum sistema de classificação é necessário usar **conjunto de teste** com dados para os quais já se conhece a classe.

Exemplo:

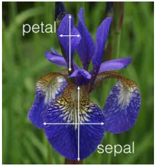
- Imagine que um botânico está interessado em distinguir 3 espécies de lírios (*iris* em inglês):

Iris Setosa

Iris Versicolor

Iris Virginica

- O botânico tirou medidas dos comprimentos e larguras da pétalas e sépalas de várias flores (**características**), e pretende classificar os lírios numa das três espécies baseado nas medições efectuadas.



- Adicionalmente, existe medidas de pétalas e sépalas de flores de lírio, para as quais se sabe a espécie (**conjunto de treino**).
- O objectivo é prever a espécie das flores medidas pelo botânico, baseado nas medidas dos lírios previamente etiquetados.

Exemplo:

- Conhecer os dados:
  Importar dados de `scikit-learn`
  ```
  >>> from sklearn import datasets
  ```
  Carregar os dados do dataset "Iris"
  ```
  >>> Iris=datasets.load_iris()
  ```
  `Iris`: variável do tipo `dictionary`, com vários campos:
  ```
  >>> Iris.keys() # ver os campos do dicionário
  ['target_names', 'data', 'target', 'DESCR', 'feature_names']
  ```
- Dados – `X` é um `np.array` de `(150,4)`:
  ```
  >>> X=Iris.data
  ```
- Classe dos dados – `trueClass` é um `np.array` de `(150,)`:
  ```
  >>> trueClass=Iris.target
  ```

Exemplo:

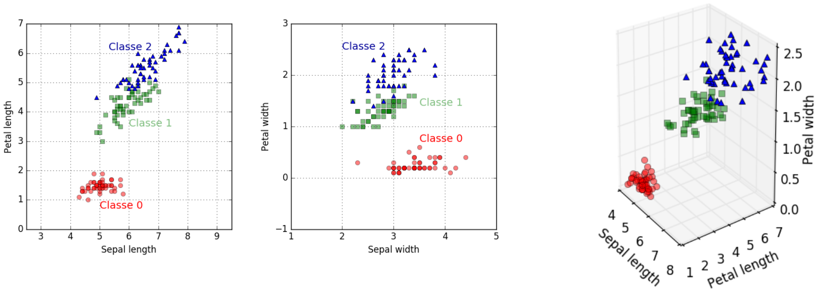
- Descrição do dataset:

```
>>> print iris.DESC
```

```
Iris Plants Database Characteristics:
    :Number of Instances: 150 (50 in each of three classes)
    :Number of Attributes: 4 numeric, predictive attributes and the class
    :Attribute Information:
        - sepal length in cm
        - sepal width in cm
        - petal length in cm
        - petal width in cm
        - class:
                - Iris-Setosa
                - Iris-Versicolour
                - Iris-Virginica
    :Summary Statistics:
    ============== ==== ==== ======= ===== ====================
                    Min  Max   Mean    SD   Class Correlation
    ============== ==== ==== ======= ===== ====================
    sepal length:   4.3  7.9   5.84   0.83    0.7826
    sepal width:    2.0  4.4   3.05   0.43   -0.4194
    petal length:   1.0  6.9   3.76   1.76    0.9490  (high!)
    petal width:    0.1  2.5   1.20   0.76    0.9565  (high!)
    ============== ==== ==== ======= ===== ====================
    :Missing Attribute Values: None
    :Class Distribution: 33.3% for each of 3 classes.
    :Creator: R.A. Fisher, July, 1988
```

Exemplo:

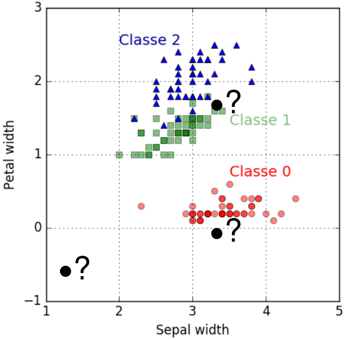
- Visualização dos dados:



- Os pontos da classe 0 (*iris setosa*) estão agrupados numa nuvem compacta e separada dos pontos das outras duas classes.
- As classes 1 e 2 têm uma zona de sobreposição de pontos de ambas as classes.

Exemplo:

- Como classificar novos dados?



Sistemas de Classificação

- Em aprendizagem automática, classificação é o problema de identificar a qual de um grupo pré-definido de classes pertence uma nova observação. A construção do modelo de classificação é baseada num conjunto de dados para as quais se conhece a classe, e é composta pelas seguinte etapas:
  1. Escolher/projectar o modelo de classificação.
  2. Treinar o modelo.
  3. Avaliar o modelo.
- Este é um problema de **aprendizagem supervisionada** (através de exemplos) - os classificadores são treinados com exemplos para os quais já se sabe a classe.

Sistemas de Classificação

Tipos de Classificação:

- Classificação **Multi-Classe**:

  Este é o cenário mais comum no contexto de classificação. Cada observação pertence a uma de um conjunto de classes. As classes são mutuamente exclusivas: uma observação não pode pertencer a mais de uma classe.

- Classificação **Binária**:

  A classificação binária é o caso da classificação multi-classe, só com duas classes.

  É importante distinguir o caso de haver duas só classes porque:

  - Existe medidas específicas de desempenho para o caso binário.
  - Problemas de detecção e recolha de informação podem ser considerados problemas de classificação binária.
  - O problema de classificação multi-label pode ser decomposto em vários problemas de classificação binária.

- Classificação **Multi-Label**:

  Na classificação multi-label existem várias classes tal como no caso de multi-classe, mas neste caso as classes não são mutuamente exclusivas. No contexto de multi-label, as classes podem ser consideradas como etiquetas e uma dada observação pode ser caracterizada por uma ou mais etiquetas. Este problema é também conhecido como *auto-tagging*.

Sistemas de Classificação

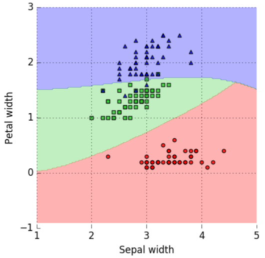
Enquadramento Teórico e Notação:
(classificação multi-classe e binária)

- Dados são representados por vetores d-dimensionais: $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}$
  Para referir os dados também se usam os termos:
  • pontos • vetores • observações • instâncias • padrões
- Cada vector de características pertence a uma única classe de um conjunto de c classes: $\Omega = \{\varpi_1, \varpi_2, \cdots, \varpi_c\}$.
  Notação:
  - $\mathbf{x} \in \varpi_\mathbf{k} \quad \Longrightarrow$ o vector pertence à classe k
  - $\mathbf{x} \in \hat{\varpi}_\mathbf{k} \quad \Longrightarrow$ o vector foi classificado na classe k
- O processo de classificação é equivalente a dividir o espaço de características num conjunto de c **regiões de decisão**.
- O processo de classificação é equivale igualmente a definir um conjunto de c **funções discriminantes**.

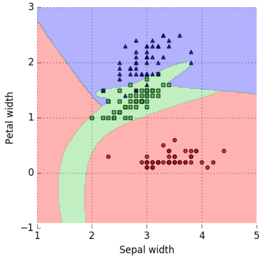
Regiões de Decisão

O processo de classificação equivale a escolher *c* regiões de decisão – tantas quanto o número de classes.

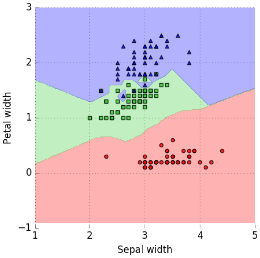
- As regiões não têm que ser contíguas e podem abranger várias zonas distintas do espaço de característica.
- A cada região é associada uma classe.
- Classificar um novo ponto corresponde a determinar em qual região esse ponto está localizado e associar-lo à classe da região.



Classificador 1



Classificador 2

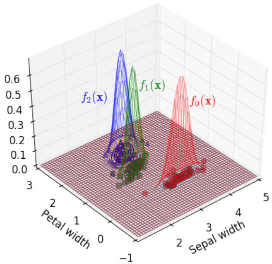


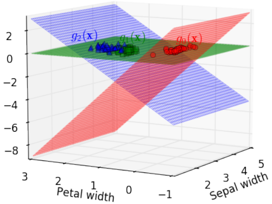
Classificador 3

Funções Discriminantes

Enumerar as regiões de decisão pode ser um processo demasiado complexo, especialmente para espaços de características de alta dimensão. Habitualmente, é preferível realizar a classificação através de **funções discriminantes**.

- Necessário definir c funções discriminantes – tantas quanto classes.
  $\mathcal{F} = \{f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_c(\mathbf{x})\}$
- A cada função é associada uma classe.
- Classificação de um novo ponto, **x**, corresponde a determinar qual das funções obteve o maior valor para **x**.
  $\mathbf{x} \in \hat{\varpi}_k$ se e só se $k = \underset{i=1,2,\ldots,c}{\operatorname{argmax}} (f_i(\mathbf{x}))$





Funções Discriminantes

O conjunto de c funções discriminantes dispensa a classificação de todos os pontos do espaço de características.

- Um conjunto de funções discriminantes $\mathcal{F} = \{f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_c(\mathbf{x})\}$ pode ser convertido noutro conjunto de funções equivalentes, transformando-as por uma função real, monótona crescente.
  - $h(\cdot) \Longrightarrow$ função real, monótona crescente.
  - $\mathcal{G} = \{g_1(\mathbf{x}), g_2(\mathbf{x}), \ldots, g_c(\mathbf{x})\}$ com $g_i(\mathbf{x}) = h(f_i(\mathbf{x}))$
  - Os dois conjuntos de funções discriminantes $\mathcal{F}$ e $\mathcal{G}$ são equivalentes (obtêm os mesmos resultados).

Exemplo: Um classificador é definido por o seguinte conjunto de funções discriminantes: $\mathcal{F} = \left\{f_1(x) = \exp(-x), f_2(x) = \exp(-x^2 + 2), f_3(x) = \exp(x/2 + 1/2)\right\}$. Pretende-se determinar as regiões de decisão do classificador.
R: Mais fácil transformar o conjunto de funções $\mathcal{F}$ pela função $h(x) = \log(x)$. Assim temos outro conjunto equivalente de funções discriminantes, mais facilmente manipulável: $\mathcal{G} = \left\{g_1(x) = -x, g_2(x) = -x^2 + 2, g_3(x) = x/2 + 1/2\right\}$.

Funções Discriminantes

O conjunto de c funções discriminantes dispensa a classificação de todos os pontos do espaço de características.

- Um conjunto de funções discriminantes $\mathcal{F} = \{f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_c(\mathbf{x})\}$ pode ser convertido noutro conjunto de funções equivalentes, transformando-as por uma função real, monótona crescente.
  - $h(\cdot) \Longrightarrow$ função real, monótona crescente.
  - $\mathcal{G} = \{g_1(\mathbf{x}), g_2(\mathbf{x}), \ldots, g_c(\mathbf{x})\}$ com $g_i(\mathbf{x}) = h(f_i(\mathbf{x}))$
  - Os dois conjuntos de funções discriminantes $\mathcal{F}$ e $\mathcal{G}$ são equivalentes (obtêm os mesmos resultados).
- Funções discriminantes também podem ser denominadas funções de lucro – procura-se aquela de retorne o maior valor.
- Pode-se igualmente realizar a classificação utilizado um conjunto de **funções de custo**. É igualmente necessário definir c funções (tantas quantas classes), mas a classificação corresponde a determinar qual função obteve o **menor** valor.
- Um conjunto $\mathcal{G}$ de funções de custo pode ser facilmente convertido noutro conjunto, $\mathcal{F}$, de funções de lucro, multiplicado cada função de custo por -1.
  $\mathcal{F} = -\mathcal{G} = \{-g_1(\mathbf{x}), -g_2(\mathbf{x}), \ldots, -g_c(\mathbf{x})\}$

Avaliação do Classificador

Para avaliar o desempenho de um classificador é necessário saber qual a **probabilidade total de errar** (ou acertar) independentemente das classes, mas também é necessário conhecer qual a probabilidade de errar e a distribuição dos erros em cada classe. Para representar a distribuição dos erros por classe usa-se uma **matriz de confusão**, que permite visualizar o desempenho do classificador.

Avaliação do Classificador

Matriz de Confusão:

- Matriz **quadrada**, **P** de $c \times c$, onde c é o número total de classes.

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1c} \\ p_{21} & p_{22} & \cdots & p_{2c} \\ \vdots & & \ddots & \\ p_{c1} & p_{c2} & \cdots & p_{cc} \end{bmatrix}$$

- Coeficientes da matriz são valores de probabilidades. $p_{ij} = p(\mathbf{x} \in \hat{\varpi}_j | \mathbf{x} \in \varpi_i)$ é a probabilidade do padrão **x** pertencer à classe ϖ_i e ser classificado na classe ϖ_j.
- Linhas da matriz referentes aos dados de uma única classe. Na primeira estão as probabilidades de acerto e de erro da classe, ϖ_1, na segunda linha da classe ϖ_2, e por aí em diante.
- A soma dos coeficientes de uma linha é igual a 1: $\sum_{i=1}^{c} p_{ki} = 1$ (corresponde à probabilidade de **x** pertencer à classe ϖ_k e de ser classificado numa das c classe - o que é o acontecimento garantido)
- No caso ideal, **P** é a matriz de identidade (não há erros).

Avaliação do Classificador

Matriz de Confusão:

- Matriz **quadrada**, **P** de $c \times c$, onde c é o número total de classes.

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1c} \\ p_{21} & p_{22} & \cdots & p_{2c} \\ \vdots & & \ddots & \\ p_{c1} & p_{c2} & \cdots & p_{cc} \end{bmatrix}$$

- Coeficientes da matriz são valores de probabilidades.
  $p_{ij} = p(\mathbf{x} \in \hat{\varpi}_j | \mathbf{x} \in \varpi_i)$ é a probabilidade do padrão **x** pertencer à classe ϖ_i e ser classificado na classe ϖ_j.
- Para calcular analiticamente o valor do coeficiente p_{ij} é necessário:
  - Conhecer as funções de densidade de probabilidade da classe ϖ_i.
    $p(\mathbf{x}|\varpi_i)$: função de densidade de **probabilidade condicionada** à classe ϖ_i.
  - Conhecer a região de decisão $\mathcal{S}_j$ da classe ϖ_j.
  - Calcular o integral $p_{ij} = \int_{\mathcal{S}_j} p(\mathbf{x}|\varpi_i) d\mathbf{x}$

Avaliação do Classificador

Matriz de Confusão e Probabilidade Total de Erro:

- A matriz de confusão representa a distribuição dos erros por classe. Para calcular **a probabilidade total do erro**, é necessário ter em conta a **probabilidade a priori** das classes, $p(\varpi_i)$ com $i = 1, \ldots, c$.
  - Probabilidade de erro da classe $\varpi_i = \sum_{j \neq i}^{c} p_{ij} = 1 - p_{ii}$
  - **Probabilidade total de erro** $= \sum_{i=1}^{c} p(\varpi_i) \left(\sum_{j \neq i}^{c} p_{ij} \right) = \sum_{i=1}^{c} p(\varpi_i)(1 - p_{ii})$
- A probabilidade total de erro é a soma dos erros de cada classe pesados pelas probabilidades a priori das classes (pela percentagem de pontos de cada classe).

Avaliação do Classificador

Questões Práticas:

- Em problemas reais, não existem funções de densidade de probabilidade condicionadas às classes, $p(\mathbf{x}|\varpi_i)$.
- Tipicamente não se sabe as regiões de decisão das classes.
- Mesmo sabendo as funções de densidade condicionada e as regiões de decisão das classes, normalmente o cálculo da probabilidade $p_{ij} = \int_{\mathcal{S}_j} p(\mathbf{x}|\varpi_i)d\mathbf{x}$ é demasiado complexo para poder ser efectuado.

SOLUÇÃO:

- Estimar as probabilidades p_{ij} e a probabilidade total de erro através de contagens de resultados de classificação de observações para as quais se sabe a classe. O conjunto de exemplos usado para a avaliação do classificador é denominado **conjunto de teste**, e deve conter exemplos diferentes dos usados para treinar o classificador.É necessário usar novos exemplos no processo de avaliação para ter uma estimativa fiável do desempenho do classificar, e medir a sua **capacidade de generalização**.

- Baseado nos exemplos classificados do conjunto de testes, os coeficientes da matriz de confusão podem ser estimados segundo: $p_{ij} = \dfrac{n_{ij}}{n_i}$

  n_{ij} número de exemplos da classe ϖ_i classificados na classe ϖ_j
  n_i número de exemplos na classe ϖ_i

Avaliação do Classificador

Exemplo: Considere um classificador definido pelo seguinte conjunto de funções discriminantes: $f_1(x) = \exp(-x)$, $f_2(x) = \exp(-x^2 + 2)$, $f_3(x) = \exp(x/2 + 1/2)$.

Baseado na seguinte tabela, determine a probabilidade total de erro e a matriz de confusão.

x	-1.5	0.5	-0.2	2.3	-2.1	2.5	1.5	-1.1	1.6	1.1	0.9	-0.1
ϖ	ϖ_2	ϖ_2	ϖ_2	ϖ_3	ϖ_1	ϖ_2	ϖ_2	ϖ_1	ϖ_3	ϖ_3	ϖ_2	ϖ_1

Avaliação do Classificador

Exemplo: Considere um classificador definido pelo seguinte conjunto de funções discriminantes: $f_1(x) = \exp(-x)$, $f_2(x) = \exp(-x^2 + 2)$, $f_3(x) = \exp(x/2 + 1/2)$.

Baseado na seguinte tabela, determine a probabilidade total de erro e a matriz de confusão.

x	-1.5	0.5	-0.2	2.3	-2.1	2.5	1.5	-1.1	1.6	1.1	0.9	-0.1
ϖ	ϖ_2	ϖ_2	ϖ_2	ϖ_3	ϖ_1	ϖ_2	ϖ_2	ϖ_1	ϖ_3	ϖ_3	ϖ_2	ϖ_1

R:

i. Aplicar a função logaritmo a todas as funções discriminantes para simplificar-las.
ii. Determinar as regiões de decisão.
iii. Classificar os dados da tabela.
iv. Determinar a probabilidade de erro e matriz de confusão.

Avaliação do Classificador

Exemplo: Considere um classificador definido pelo seguinte conjunto de funções discriminantes: $f_1(x) = \exp(-x)$, $f_2(x) = \exp(-x^2 + 2)$, $f_3(x) = \exp(x/2 + 1/2)$.

Baseado na seguinte tabela, determine a probabilidade total de erro e a matriz de confusão.

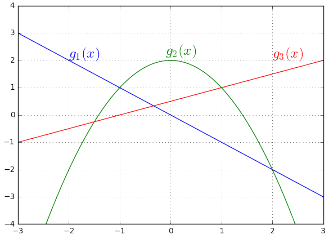
x	-1.5	0.5	-0.2	2.3	-2.1	2.5	1.5	-1.1	1.6	1.1	0.9	-0.1
ϖ	ϖ_2	ϖ_2	ϖ_2	ϖ_3	ϖ_1	ϖ_2	ϖ_2	ϖ_1	ϖ_3	ϖ_3	ϖ_2	ϖ_1

R:

i. $g_i(x) = \ln(f_i(x))$

$g_1(x) = -x$, $g_2(x) = -x^2 + 2$, $g_3(x) = x/2 + 1/2$.

ii. $\mathcal{S}_1 =]-\infty, -1]$, $\quad \mathcal{S}_2 = [-1, +1]$, $\quad \mathcal{S}_3 = [+1, +\infty[$



Avaliação do Classificador

Exemplo: Considere um classificador definido pelo seguinte conjunto de funções discriminantes: $f_1(x) = \exp(-x)$, $f_2(x) = \exp(-x^2 + 2)$, $f_3(x) = \exp(x/2 + 1/2)$.

Baseado na seguinte tabela, determine a probabilidade total de erro e a matriz de confusão.

x	-1.5	0.5	-0.2	2.3	-2.1	2.5	1.5	-1.1	1.6	1.1	0.9	-0.1
ϖ	ϖ_2	ϖ_2	ϖ_2	ϖ_3	ϖ_1	ϖ_2	ϖ_2	ϖ_1	ϖ_3	ϖ_3	ϖ_2	ϖ_1
$\hat{\varpi}$	ϖ_1	ϖ_2	ϖ_2	ϖ_3	ϖ_1	ϖ_3	ϖ_3	ϖ_1	ϖ_3	ϖ_3	ϖ_2	ϖ_2
	X					**X**	**X**					**X**

R:

iii. Resultados errados de classificação marcados com um "**X**"

- $N = 12$ número total de pontos classificados
- Classe ϖ_1: $n_1 = 3 \Longrightarrow \quad p(\varpi_1) = \frac{3}{12} = \frac{1}{4}$
  $n_{11} = 2 \quad n_{12} = 1 \quad n_{13} = 0$
- Classe ϖ_2: $n_2 = 6 \Longrightarrow \quad p(\varpi_2) = \frac{6}{12} = \frac{1}{2}$
  $n_{21} = 1 \quad n_{22} = 3 \quad n_{23} = 2$
- Classe ϖ_3: $n_3 = 3 \Longrightarrow \quad p(\varpi_3) = \frac{3}{12} = \frac{1}{4}$
  $n_{31} = 0 \quad n_{32} = 0 \quad n_{33} = 3$

Avaliação do Classificador

Exemplo: Considere um classificador definido pelo seguinte conjunto de funções discriminantes: $f_1(x) = \exp(-x)$, $f_2(x) = \exp(-x^2 + 2)$, $f_3(x) = \exp(x/2 + 1/2)$.

Baseado na seguinte tabela, determine a probabilidade total de erro e a matriz de confusão.

x	-1.5	0.5	-0.2	2.3	-2.1	2.5	1.5	-1.1	1.6	1.1	0.9	-0.1
ϖ	ϖ_2	ϖ_2	ϖ_2	ϖ_3	ϖ_1	ϖ_2	ϖ_2	ϖ_1	ϖ_3	ϖ_3	ϖ_2	ϖ_1
$\hat{\varpi}$	ϖ_1	ϖ_2	ϖ_2	ϖ_3	ϖ_1	ϖ_3	ϖ_3	ϖ_1	ϖ_3	ϖ_3	ϖ_2	ϖ_2
	X					**X**	**X**					**X**

R:

iv. Matriz de confusão e probabilidade total de erro

- Classe ϖ_1: $n_1 = 3$ e $n_{11} = 2 \quad n_{12} = 1 \quad n_{13} = 0$
  $p_{11} = \frac{n_{11}}{n_1} = \frac{2}{3} \quad p_{12} = \frac{1}{3} \quad p_{13} = 0$
- Classe ϖ_2: $n_2 = 6$ e $n_{21} = 1 \quad n_{22} = 3 \quad n_{23} = 2$
  $p_{21} = \frac{1}{6} \quad p_{22} = \frac{1}{2} \quad p_{23} = \frac{1}{3}$
- Classe ϖ_3: $n_3 = 3$ e $n_{31} = 0 \quad n_{32} = 0 \quad n_{33} = 3$
  $p_{31} = \quad p_{32} = 0 \quad p_{33} = 1$

$$\mathbf{P} = \begin{bmatrix} \frac{2}{3} & \frac{1}{3} & 0 \\ \frac{1}{6} & \frac{1}{2} & \frac{1}{3} \\ 0 & 0 & 1 \end{bmatrix}$$

Probabilidade total de erro $= \sum_{i=1}^{3} (1 - p_{ii})\, p(\varpi_i) = \left(1 - \frac{2}{3}\right)\frac{3}{12} + \left(1 - \frac{1}{2}\right)\frac{6}{12} + (1-1)\frac{3}{12} = \frac{4}{12} = \frac{1}{3}$

Pode-se calcular diretamente:
Há 4 erros em 12 exemplos $\Longrightarrow$ probabilidade total de erro $= \frac{4}{12}$

Avaliação do Classificador

Matriz de Confusão Não-Normalizada:

- Normalmente, os coeficientes $p_{ij} = p(\mathbf{x} \in \varpi_j | \mathbf{x} \in \varpi_i)$ da matriz de confusão são estimados através de contagens dos resultados de classificação de dados previamente classificados (conjunto de teste).
- Por vezes é mais prático não dividir as contagens pelo número total de pontos da classe. Assim a soma dos valores das linha da matriz de confusão passa a ser o número total de exemplos em cada classe.
- n_{ij} Número de exemplos da classe ϖ_i classificados na classe ϖ_j

$$\text{Matriz de Confusão Não Normalizada} = \begin{bmatrix} n_{11} & n_{12} & \cdots & n_{1c} \\ n_{21} & n_{22} & \cdots & n_{2c} \\ \vdots & & \ddots & \\ n_{c1} & n_{c2} & \cdots & n_{cc} \end{bmatrix}$$

<u>Nota:</u> com esta matriz também se pode calcular a probabilidade total do erro.

Exemplo: Conjunto de dados `Iris`

- Dados: flores de lírio representadas com quatro características (pontos a 4 dimensões $\mathbf{x} = [x_1, x_2, x_3, x_4]^\top$).
- Classes: 3 espécies de lírio - setosa, versicolor, virginica (classe $\varpi_1, \varpi_2,$ e ϖ_3 respetivamente).
- 150 observações, 50 de cada classe

Avaliação do Classificador

Matriz de Confusão Não-Normalizada:

- Normalmente, os coeficientes $p_{ij} = p(\mathbf{x} \in \varpi_j | \mathbf{x} \in \varpi_i)$ da matriz de confusão são estimados através de contagens dos resultados de classificação de dados previamente classificados (conjunto de teste).
- Por vezes é mais prático não dividir as contagens pelo número total de pontos da classe. Assim a soma dos valores das linha da matriz de confusão passa a ser o número total de exemplos em cada classe.
- n_{ij} Número de exemplos da classe ϖ_i classificados na classe ϖ_j

  Matriz de Confusão Não Normalizada$=\begin{bmatrix} n_{11} & n_{12} & \cdots & n_{1c} \\ n_{21} & n_{22} & \cdots & n_{2c} \\ \vdots & & \ddots & \\ n_{c1} & n_{c2} & \cdots & n_{cc} \end{bmatrix}$

  Nota: com esta matriz também se pode calcular a probabilidade total do erro.

Exemplo: Conjunto de dados `Iris`

Resultados da classificação com o seguinte conjunto de funções discriminantes:

$$\begin{bmatrix} f_1(\mathbf{x}) \\ f_2(\mathbf{x}) \\ f_3(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} -8 & 1 & 5 & -4 & -1 \\ 20 & 0 & -9 & 4 & -9 \\ -25 & 0 & 3 & 0 & 11 \end{bmatrix} \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \qquad \mathbf{P} = \begin{bmatrix} 50 & 0 & 0 \\ 0 & 32 & 18 \\ 0 & 5 & 45 \end{bmatrix}$$

Probabilidade total de erro $= (18 + 5)/150 \approx 15.3\%$

Avaliação do Classificador

Classificação Binária

É importante analisar o caso particular da classificação em 2 classes porque esta surge em diversos diversos domínios de aplicação.

- Sistemas de detecção - um paciente tem ou não uma doença?
- Sistemas de alarme - existe ou não uma intrusão?
- Sistemas de identificação - é ou não a pessoa correcta?
- Sistemas de etiquetação - uma música tem ou não instrumentos de cordas?
- Sistemas de recolha de informação - a pesquisa retornou ou não o pretendido?

- É habitual referir as duas classes como positivos e negativos $(\varpi_{\mathrm{p}}, \varpi_{\mathrm{n}})$. Tipicamente a classe dos positivos representa a existência ou detecção de uma dada condição, situação, registo, teste, etc.
- Em muitos casos práticos, o número de exemplos positivos é significativamente menor que os negativo. Nestas situações, a probabilidade total de erro (ou acertos) não é uma boa medida de desempenho. Sistemas que classifiquem tudo como negativo obtêm uma percentagem de acertos elevada.
- Para o caso da classificação binária, existem várias outras métricas de desempenho que se adequam melhor a esta situação.

Avaliação do Classificador

Classificação Binária - Métricas de Desempenho

As métricas de desempenho para o caso da classificação binária, todas elas baseada nos valores matriz de confusão não-normalizada (para duas classes, a matriz de confusão também se denomina “tabela de contingência”).

	$\hat{\varpi}_{\mathrm{p}}$	$\hat{\varpi}_{\mathrm{n}}$
ϖ_{p}	**T**rue **P**ositives	**F**alse **N**egatives
ϖ_{n}	**F**alse **P**ositives	**T**rue **N**egatives

- Classe dos positivos ϖ_{p}:
  - Número de exemplos: TP+FN
  - $p(\varpi_{\mathrm{p}}) = \dfrac{\mathrm{TP} + \mathrm{FN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP} + \mathrm{TN}}$
- Classe dos negativos ϖ_{n}:
  - Número de exemplos: FP+TN
  - $p(\varpi_{\mathrm{n}}) = \dfrac{\mathrm{FP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP} + \mathrm{TN}}$

- Na classificação binária existem várias métricas de desempenho que refletem diferentes especificidades do desempenho do classificador. A escolha de quais métricas se devem usar depende da área de aplicação e da importância e da proporção dos dois tipos de erros possíveis. Por exemplo em diagnósticos médicos, um falso positivo (detetar uma doença quando esta não existe) tem um custo diferente de um falso negativo (não detetar uma doença quando esta existe) e em medicina é comum usar o *recall* em conjunção com a *sensitivity*. Em problemas de recolha de informação é habitual usar o *recall* e *precision*.

Avaliação do Classificador

Classificação Binária - Métricas de Desempenho

As métricas de desempenho para o caso da classificação binária, todas elas baseada nos valores matriz de confusão não-normalizada (para duas classes, a matriz de confusão também se denomina “tabela de contingência”).

	$\hat{\varpi}_{\mathrm{p}}$	$\hat{\varpi}_{\mathrm{n}}$
ϖ_{p}	**T**rue **P**ositives	**F**alse **N**egatives
ϖ_{n}	**F**alse **P**ositives	**T**rue **N**egatives

- Classe dos positivos ϖ_{p}:
  - Número de exemplos: TP+FN
  - $p\left(\varpi_{\mathrm{p}}\right) = \dfrac{\mathrm{TP} + \mathrm{FN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP} + \mathrm{TN}}$
- Classe dos negativos ϖ_{n}:
  - Número de exemplos: FP+TN
  - $p\left(\varpi_{\mathrm{n}}\right) = \dfrac{\mathrm{FP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP} + \mathrm{TN}}$

- Há oito métricas básicas que se podem calcular da matriz de confusão. Estas são obtidas dividindo os quatros resultados da tabela pela soma das linhas ou das colunas.
- Ao somar nas linhas, estamos a ter em conta percentagens de acertos ou erros relativos ao número total de pontos de cada classe. Estes valores não são afectados por haver mais ou menos exemplos de uma das classes.
- Ao somar nas colunas estamos a ter em conta percentagens de acertos ou erros relativos ao número total de pontos classificados em cada classe. Estes valores são afectados pela proporção entre o número de exemplos em cada classe.

Avaliação do Classificador

Classificação Binária - Métricas de Desempenho

	$\hat{\varpi}_\mathrm{p}$	$\hat{\varpi}_\mathrm{n}$
ϖ_p	**T**rue **P**ositives	**F**alse **N**egatives
ϖ_n	**F**alse **P**ositives	**T**rue **N**egatives

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} = \begin{bmatrix} \frac{\mathrm{TP}}{\mathrm{TP+FN}} & \frac{\mathrm{FN}}{\mathrm{TP+FN}} \\ \frac{\mathrm{FP}}{\mathrm{FP+TN}} & \frac{\mathrm{TN}}{\mathrm{FP+TN}} \end{bmatrix}$$

(matriz de confusão normalizada)

$$\text{Probabilidade Total de Erro} = \frac{\mathrm{FP+FN}}{\mathrm{TP+FP+TN+FN}}$$

- TP-rate $= \dfrac{\mathrm{TP}}{\mathrm{TP+FN}} = p_{11}$
  porção dos positivos bem classificados
  Sinónimos: • **recall** • sensitivity
- FN-rate $\dfrac{\mathrm{FN}}{\mathrm{TP+FN}} = p_{12}$
  porção dos positivos mal classificados
- TN rate $= \dfrac{\mathrm{TN}}{\mathrm{FP+TN}} = p_{22}$
  porção dos negativos bem classificados
  Sinónimos: • specificity
- FP rate $= \dfrac{\mathrm{FP}}{\mathrm{FP+TN}} = p_{21}$
  porção dos negativos mal classificados
  Sinónimos: • false alarm • fall-out

Avaliação do Classificador

Classificação Binária - Métricas de Desempenho

	$\hat{\varpi}_p$	$\hat{\varpi}_n$
ϖ_p	**T**rue **P**ositives	**F**alse **N**egatives
ϖ_n	**F**alse **P**ositives	**T**rue **N**egatives

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} = \begin{bmatrix} \frac{TP}{TP+FN} & \frac{FN}{TP+FN} \\ \frac{FP}{FP+TN} & \frac{TN}{FP+TN} \end{bmatrix}$$

(matriz de confusão normalizada)

$$\text{Probabilidade Total de Erro} = \frac{FP+FN}{TP+FP+TN+FN}$$

- PPV Positive Predicted Value $= \dfrac{TP}{TP+FP}$
  porção dos classificados como positivos bem classificados
  Sinónimos: • **precision**
- FDR False Dicovery Rate $= \dfrac{FP}{TP+FP}$
  porção dos classificados como positivos mal classificados
  Notar que $FDR = 1 - PPV$
- NPV Negative Predicted Value $= \dfrac{TN}{TN+FN}$
  porção dos classificados como negativos bem classificados
- FOR False Omission Rate $= \dfrac{FN}{TN+FN}$
  porção dos classificados como negativos mal classificados

Avaliação do Classificador

Classificação Binária - Métricas de Desempenho

	$\hat{\varpi}_p$	$\hat{\varpi}_n$
ϖ_p	**T**rue **P**ositives	**F**alse **N**egatives
ϖ_n	**F**alse **P**ositives	**T**rue **N**egatives

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} = \begin{bmatrix} \frac{TP}{TP+FN} & \frac{FN}{TP+FN} \\ \frac{FP}{FP+TN} & \frac{TN}{FP+TN} \end{bmatrix}$$

(matriz de confusão normalizada)

$$\text{Probabilidade Total de Erro} = \frac{FP+FN}{TP+FP+TN+FN}$$

- Precision + recall são as métricas preferidas em problemas de recolha de informação. Porém, classificadores triviais podem obter um desempenho elevado ou na precision ou no recall. Só classificadores válidos obtêm um desempenho alto em ambas as métricas.
- Métricas derivadas da precision e do recall.
  - $\text{F-Score} = \dfrac{1}{1/\text{recall} + 1/\text{precision}} = 2 \times \dfrac{\text{precision} \times \text{recall}}{\text{precision+recall}}$
    (média harmónica entre precision e recall)
  - $\text{G-Score} = \sqrt{\text{precision} \times \text{recall}}$
    (média geométrica entre precision e recall)

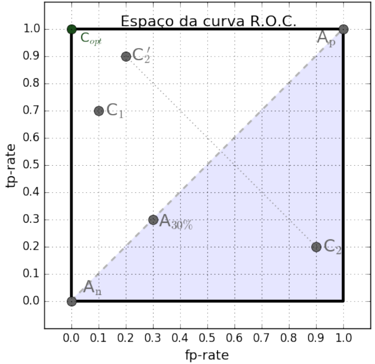
Avaliação do Classificador

Classificação Binária - Curvas ROC

- As curvas ROC (Receiving Operating Characteristics) são gráficos que servem para ilustrar e comparar o desempenho dum ou mais classificadores.
- O espaço da curva ROC é definido no eixo das abcissas por a taxa de falsos positivos (fp-rate) e no eixo das ordenadas por a taxa de verdadeiros positivos (tp-rate). Os valores estão limitados no intervalo $[0, 1]$.
- As curvas ROC são uma ferramenta importante para a escolha dos modelos ótimos e no descarte dos sub-ótimos.
- Um resultado de classificação corresponde a um único ponto na curva.
- Inerente a muitos modelos de classificação está a possibilidade de variar o limiar de decisão. Ao analisar um modelo para vários valores do limiar resulta numa curva no espaço ROC, curva essa que pode ser usada na escolher do limiar mais apropriado ao problema.

Classificação Binária - Curvas ROC

Curvas ROC para 7 resultados de classificação: $\mathrm{A_n}$, $\mathrm{A_p}$, $\mathrm{A}_{30\%}$, C_1, C_2, C_2', e C_{opt}.

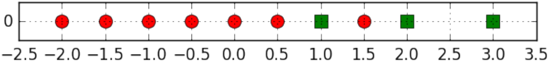


- C_{opt}: classificador perfeito (não há erros).
- Classificadores aleatórios: $\mathrm{A_n}$, $\mathrm{A_p}$ e $\mathrm{A}_{30\%}$ Todos os classificadores localizados na **linha diagonal a tracejado** de $(0,0)$ a $(1,1)$ são obtidos com escolhas aleatórias.
- Triângulo inferior: zona dos classificadores **piores** que aleatórios. Dá para reposicionar os classificadores no triângulo superior ao **inverter** a decisão de classificação (trocar os positivos pelos negativos e vice versa). Exemplo: C_2 e C_2'

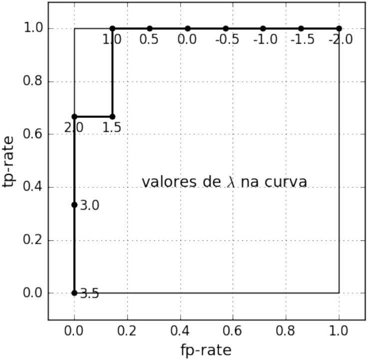
Avaliação do Classificador

Curvas ROC - Exemplo 1:

Considere o conjunto de $N = 10$ pontos 1D divididos em duas classes ($\square \in \varpi_\mathrm{p}$, $\circ \in \varpi_\mathrm{n}$).



Considere igualmente o seguinte processo de classificação: $x \in \hat{\varpi}_\mathrm{p}$ se $x \geq \lambda$, $x \in \hat{\varpi}_\mathrm{n}$ se $x < \lambda$. Ao variar o limiar λ de $-\infty$ a $+\infty$ obtemos um gráfico no espaço da curva ROC (neste exemplo bastou variar λ de 3.5 a -2 de 0.5 em 0.5 unidades).



λ	tp-rate	fp-rate
+3.5	0	0
+3.0	1/3	0
+2.5	1/3	0
+2.0	2/3	0
+1.5	2/3	1/7
+1.0	1	1/7
+0.5	1	2/7
0.0	1	3/7
−0.5	1	4/7
−1.0	1	5/7
−1.5	1	6/7
−2.0	1	1

Avaliação do Classificador

Curvas ROC - Exemplo 2:

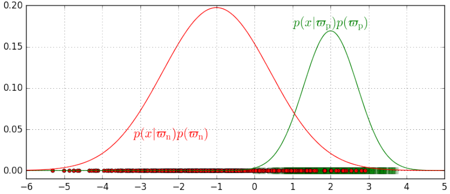
Considere o conjunto pontos 1D dividido em duas classes e distribuídos segundo:

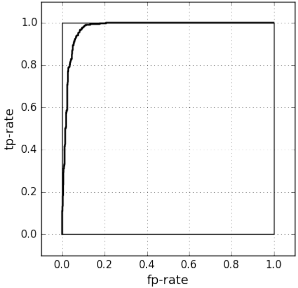
Classe dos positivos: $p\left(x|\varpi_{\mathrm{p}}\right) = \frac{1}{\sqrt{\pi}} \exp\left\{-(x-2)^2\right\}$ e $p(\varpi_{\mathrm{p}}) = 0.3$

Classe dos negativos: $p\left(x|\varpi_{\mathrm{n}}\right) = \frac{1}{\sqrt{4\pi}} \exp\left\{-\frac{1}{4}(x+1)^2\right\}$ e $p(\varpi_{\mathrm{n}}) = 0.7$

Considere igualmente o seguinte processo de classificação: $x \in \hat{\varpi}_{\mathrm{p}}$ se $x \geq \lambda, x \in \hat{\varpi}_{\mathrm{n}}$ se $x < \lambda$.

No gráfico da direita estão representados $N = 1000$ pontos e as funções de densidade das duas classes e no gráfico da esquerda a curva ROC com a variação do valor do limiar λ.





Avaliação do Classificador

Outras Curvas e Medidas de Desempenho:

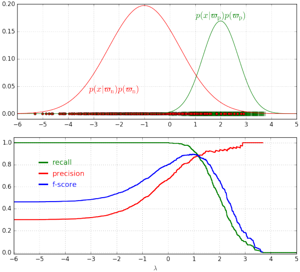
- **AUC -Area Under the ROC Curve:**
  A área da curva ROC é uma maneira de combinar as duas métricas tp-rate e fp-rate numa só. AUC pode ser interpretado como a capacidade de discriminar correctamente as observações positivas das negativas.
- **Curvas DET - Detection Error Tradeoff:**
  Esta curva serve para visualizar as taxas de erro de um classificador mostrando as detecções falhadas (fn-rate) versus falsos alarmes (fp-rate).
- **Curvas de Precision vs Recall:**
  Curvas de precision versus recall são muito usados em recolha de informação. Estas curvas têm uma característica distinta das curvas ROC ou DET. As curvas ROC e DET não são afetadas pelo número de exemplos positivos ou negativos, ao passo que as curvas precision-recall sim. Isto é: diferentes proporções de exemplos positivos e negativos produzem diferentes gráficos.

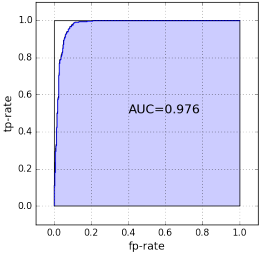
Avaliação do Classificador

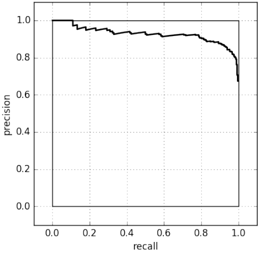
Exemplo 2:

$$p\left(x|\varpi_{\mathrm{p}}\right) = \frac{1}{\sqrt{\pi}} \exp\left\{-(x-2)^2\right\} \text{ e } p(\varpi_{\mathrm{p}}) = 0.3$$

$$p\left(x|\varpi_{\mathrm{n}}\right) = \frac{1}{\sqrt{4\pi}} \exp\left\{-\frac{1}{4}(x+1)^2\right\} \text{ e } p(\varpi_{\mathrm{n}}) = 0.7$$







Avaliação do Classificador

Exemplo 2:

$$p\left(x|\varpi_{\mathrm{p}}\right) = \frac{1}{\sqrt{\pi}} \exp\left\{-(x-2)^2\right\} \text{ e } p(\varpi_{\mathrm{p}}) = 0.1$$

$$p\left(x|\varpi_{\mathrm{n}}\right) = \frac{1}{\sqrt{4\pi}} \exp\left\{-\frac{1}{4}(x+1)^2\right\} \text{ e } p(\varpi_{\mathrm{n}}) = 0.9$$