

# Introdução ao Aprendizado de Máquina

Lucas Gonçalves de Moura Leite

# Métricas de Avaliação

# Métricas de avaliação

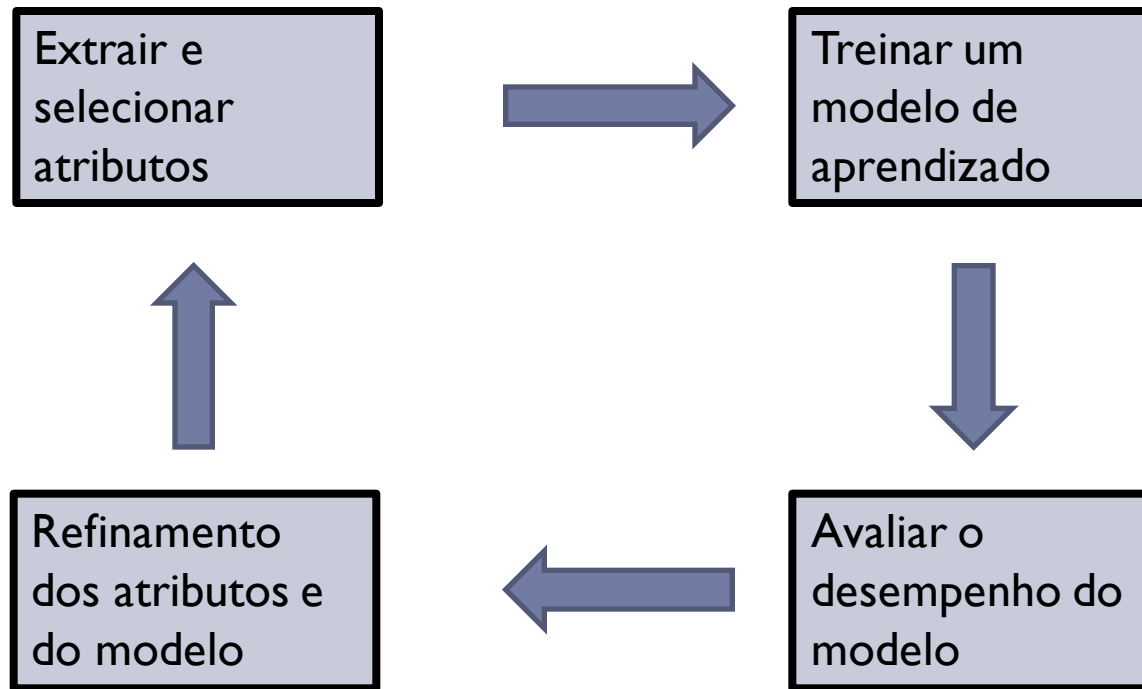
---

- ▶ Classificação
  - ▶ Acurácia
- ▶ Regressão
  - ▶  $R^2$
- ▶ Desvantagens destas métricas e alternativas
- ▶ Como escolher a métrica mais adequada



# Processo de AM

---



# Avaliação

---

- ▶ Aplicações diferentes tem objetivos diferentes
- ▶ Acurácia é bastante usada mas outras métricas podem ser usadas
  - ▶ Busca na web (satisfação do usuário)
  - ▶ Retorno financeiro (Comércio)
  - ▶ Aumento na taxa de sobrevivência (aplicações médicas)



# Avaliação para dados desbalanceados

---

- ▶ Suponha um problema com duas classes

- ▶ Relevante (R) – Classe positiva
- ▶ Irrelevante (I) – Classe negativa

- ▶ Exemplo

- ▶ Sistemas de recomendação
- ▶ Identificação de fraudes em cartões

- ▶ 1 em cada 1000 é relevante

- ▶  $Acc = \frac{\text{predições corretas}}{\text{total de itens}}$



# Avaliação para dados desbalanceados

---

- ▶ Você projetou um classificador com 99.9% de acurácia.
- ▶ Quão bom é esse resultado?



# Avaliação para dados desbalanceados

---

- ▶ Você projetou um classificador com 99.9% de acurácia.
- ▶ Quão bom é esse resultado?
- ▶ Comparar com um classificador muito simples. Ele sempre diz que um item é irrelevante.





# Avaliação para dados desbalanceados

---

- ▶ Você projetou um classificador com 99.9% de acurácia.
- ▶ Quão bom é esse resultado?
- ▶ Comparar com um classificador muito simples. Ele sempre diz que um item é irrelevante.
  - ▶ Qual a sua taxa de acerto?



# Classificador Dummy

---

- ▶ Verificador de sanidade
- ▶ Baseline para classificação
- ▶ Tipos
  - ▶ Mais frequente (`most_frequent`)
  - ▶ Aleatório (`uniform`)
  - ▶ Aleatório com distribuição igual aos dados de treino (`stratified`)
  - ▶ Constante e configurado pelo usuário (`constant`)



# Classificador = Dummy

---

## ▶ Possíveis causas

- ▶ Erros nos atributos (valores faltantes, errados ...)
- ▶ Overfitting (má escolha dos hiperparâmetros)
- ▶ Desbalanceamento



# Regressores Dummy

---

## ▶ Tipos

- ▶ Média das saídas do treinamento (mean)
- ▶ Mediana das saídas do treinamento (median)
- ▶ Quantil das saídas de treinamento. 0 para mínimo, 0.5 para média e 1 para máximo (quantile)
- ▶ Constante e configurado pelo usuário (constant)



# Matriz de Confusão

---

<u>True</u> negative	TN	FP
<u>True</u> positive	FN	TP
	<u>Predicted</u> negative	<u>Predicted</u> positive

Label 1 = positive class  
(class of interest)

Label 0 = negative class  
(everything else)

TP = true positive

FP = false positive (Type I error)

TN = true negative

FN = false negative (Type II error)

# Métricas

---

True negative	TN = 400	FP = 7	
True positive	FN = 17	TP = 26	
	Predicted negative	Predicted positive	$N = 450$

$$\triangleright Acc = \frac{TP+TN}{TP+TN+FP+FN}$$

$$\triangleright Acc = 0.95$$



# Erro de Classificação

---

True negative	TN = 400	FP = 7	
True positive	FN = 17	TP = 26	
	Predicted negative	Predicted positive	$N = 450$

$$\blacktriangleright Err = \frac{FP+FN}{TP+TN+FP+FN}$$

$$\blacktriangleright Err = 0.05$$



# Recall

True negative	TN = 400	FP = 7	
True positive	FN = 17	TP = 26	
	Predicted negative	Predicted positive	$N = 450$

$$\triangleright Rec = \frac{TP}{TP+FN}$$

$$\triangleright Rec = 0.6$$

- ▶ **Conhecido como:**
  - ▶ Sensitividade
  - ▶ True Positive Rate (Taxa de verdadeiros positivos)
- ▶ **Exemplo**
  - ▶ Detecção de Cancer



# Precision

---

True negative	TN = 400	FP = 7	
True positive	FN = 17	TP = 26	
	Predicted negative	Predicted positive	$N = 450$

$$\triangleright Pre = \frac{TP}{TP+FP}$$

$$\triangleright Pre = 0.79$$

## ▶ Exemplo

- ▶ Sistemas de recomendação



# False Positive Rate

---

True negative	TN = 400	FP = 7	
True positive	FN = 17	TP = 26	
	Predicted negative	Predicted positive	$N = 450$

$$\triangleright Pre = \frac{FP}{TN+FP}$$

$$\triangleright Pre = 0.02$$

► Conhecido como:

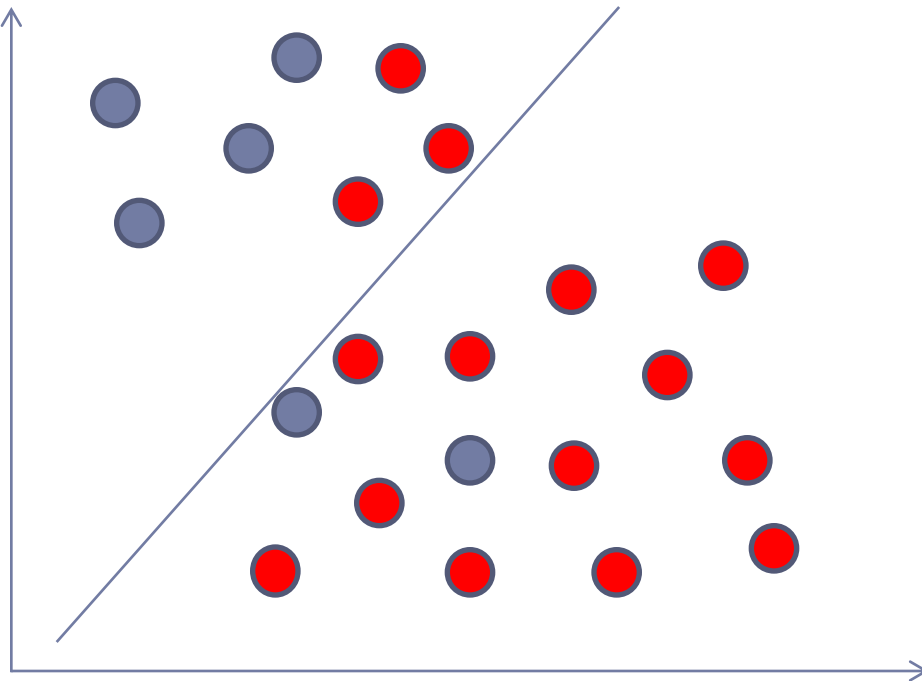
► Especificidade



# Precision x Recall

---

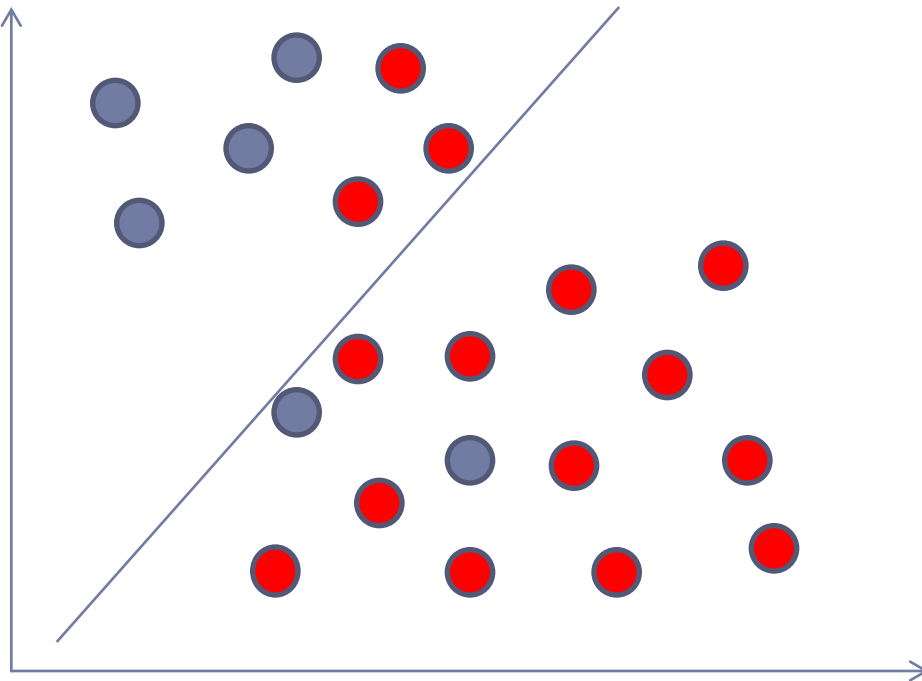
TN	FP
FN	TP



# Precision x Recall

---

TN = 12	FP = 3
FN = 2	TP = 4

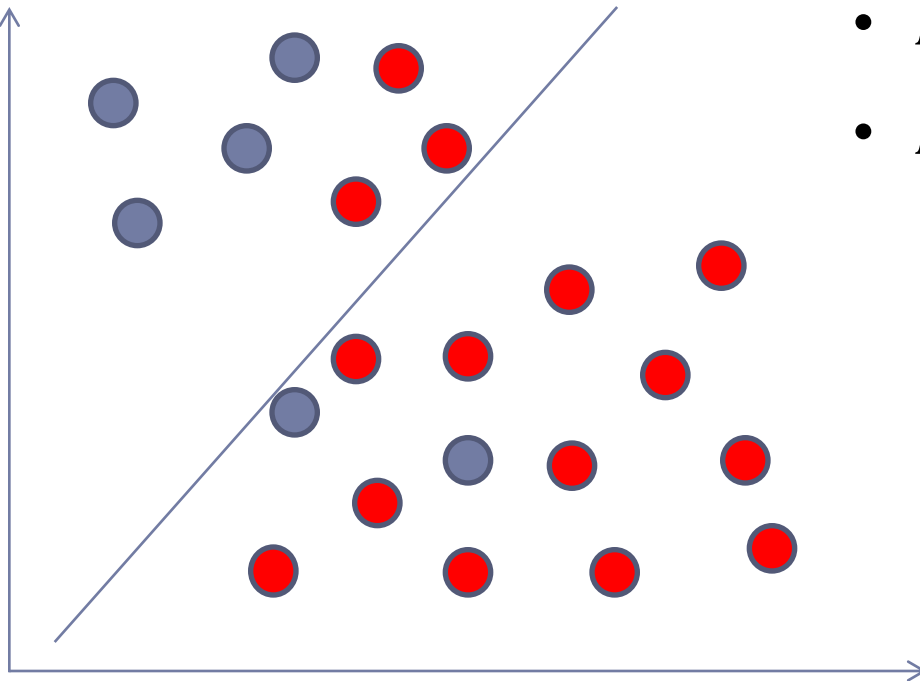


# Precision x Recall

---

TN = 12	FP = 3
FN = 2	TP = 4

- $Pre = \frac{TP}{TP+FP} = \frac{4}{4+3} = 0.57$
- $Rec = \frac{TP}{TP+FN} = \frac{4}{4+2} = 0.66$

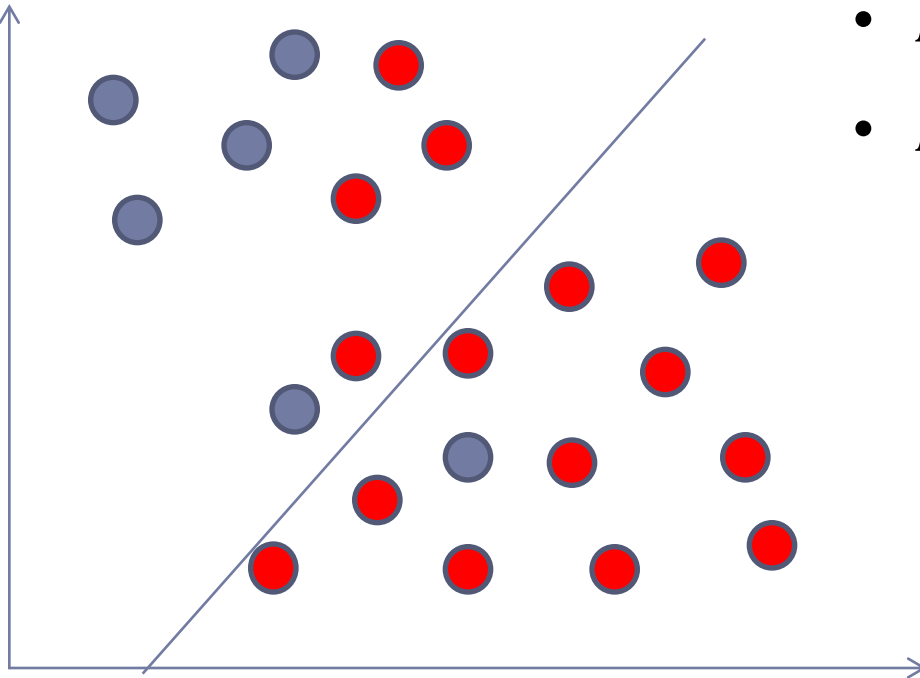


# Precision x Recall

---

TN = 11	FP = 4
FN = 1	TP = 5

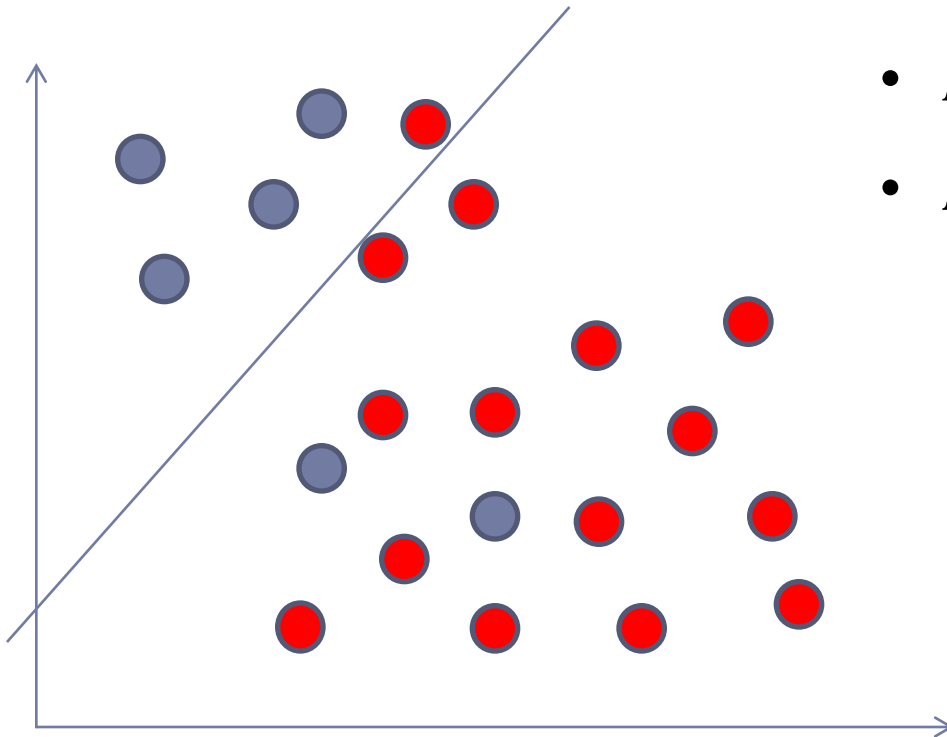
- $Pre = \frac{TP}{TP+FP} = \frac{5}{5+4} = 0.55$
- $Rec = \frac{TP}{TP+FN} = \frac{5}{5+1} = 0.83$



# Precision x Recall

---

TN = 14	FP = 1
FN = 2	TP = 4



- $Pre = \frac{TP}{TP+FP} = \frac{4}{4+1} = 0.8$
- $Rec = \frac{TP}{TP+FN} = \frac{4}{4+2} = 0.66$



# Aplicações (Precision e Recall)

---

## ▶ Recall

- ▶ Aplicações médicas
- ▶ Busca por informações
- ▶ Apoio humano para filtrar falsos positivos

## ▶ Precision

- ▶ Recomendação
- ▶ Aplicações que lidam direto com um cliente humano





# F-measure

---

- ▶ F1-score

- ▶ 
$$F1 = 2 \frac{Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FN + FP}$$



# Confiança em classificadores

# Confiança na classificação

---

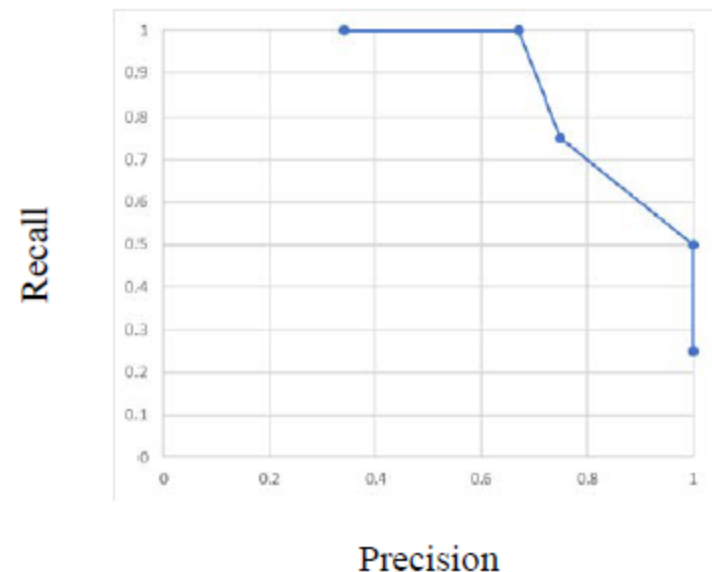
- ▶ Muitos classificadores fornecem como saída uma medida que indica a certeza da classificação
- ▶ Usualmente a classificação é feita com um limiar
- ▶ A mudança desse limiar tem como resultado uma série de classificações que formam uma curva de desempenho



# Variando o limiar

True Label	Classifier score
0	-27.6457
0	-25.8486
0	-25.1011
0	-24.1511
0	-23.1765
0	-22.575
0	-21.8271
0	-21.7226
0	-19.7361
0	-19.5768
0	-19.3071
0	-18.9077
0	-13.5411
0	-12.8594
1	-3.9128
0	-1.9798
1	1.824
0	4.74931
1	15.234624
1	21.20597

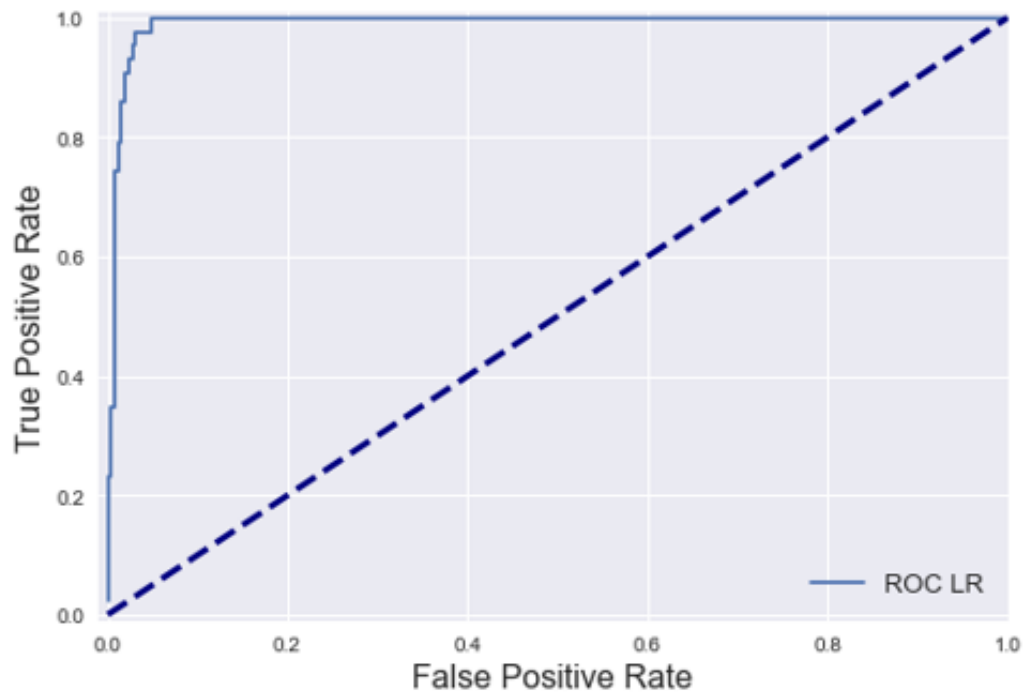
Classifier score threshold	Precision	Recall
-20	$4/12=0.34$	$4/4=1.00$
-10	$4/6=0.67$	$4/4=1.00$
0	$3/4=0.75$	$3/4=0.75$
10	$2/2=1.0$	$2/4=0.50$
20	$1/1=1.0$	$1/4 = 0.25$



Curva Precision x Recall

# Curva ROC

- ▶ Eixo x – False Positive Rate
- ▶ Eixo y – True Positive Rate



# Avaliação de classificadores com múltiplas classes

# Métricas para Classificadores Multiclasse

---

- ▶ A extensão de algumas métricas é bastante simples
  - ▶ Matriz de confusão
- ▶ Outras tem como princípio o calculo de uma média da métrica para todas as classes
  - ▶ Diferentes formas de calcular médias ponderadas



# Micro e Macro Average

---

- ▶ Formas de ponderar os resultados por classe
- ▶ Macro
  - ▶ Média dos resultados de cada classe
  - ▶ Cada classe tem o mesmo peso
- ▶ Micro
  - ▶ Computa a métrica com o resultado do classificador
  - ▶ Não separa por classe
  - ▶ Cada exemplo tem o mesmo peso





# Macro Average

---

Class	Predicted Class	Correct?
orange	lemon	0
orange	lemon	0
orange	apple	0
orange	orange	1
orange	apple	0
lemon	lemon	1
lemon	apple	0
apple	apple	1
apple	apple	1

## Macro-average:

- Each class has equal weight.
1. Compute metric within each class
  2. Average resulting metrics across classes

<u>Class</u>	<u>Precision</u>
orange	$1/5 = 0.20$
lemon	$1/2 = 0.50$
apple	$2/2 = 1.00$

Macro-average precision:  
 $(0.20 + 0.50 + 1.00) / 3 = \mathbf{0.57}$



# Micro Average

---

Class	Predicted Class	Correct?
orange	lemon	0
orange	lemon	0
orange	apple	0
orange	orange	1
orange	apple	0
lemon	lemon	1
lemon	apple	0
apple	apple	1
apple	apple	1

## Micro-average:

- Each instance has equal weight.
  - Largest classes have most influence
1. Aggregate outcomes across all classes
  2. Compute metric with aggregate outcomes

Micro-average precision:

$$4 / 9 = \mathbf{0.44}$$



# Micro x Macro

---

- ▶ Se os dados são balanceados, Micro e Macro são aproximadamente iguais
- ▶ Se dados são desbalanceados
  - ▶ Se quiser uma métrica enviesada para as classes majoritárias, use Micro.
  - ▶ Se quiser uma métrica enviesada para as classes minoritárias, use Macro.
  - ▶ Se  $\text{Micro} \ll \text{Macro}$
  - ▶ Se  $\text{Micro} \gg \text{Macro}$



# Micro x Macro

---

- ▶ Se os dados são balanceados, Micro e Macro são aproximadamente iguais
- ▶ Se dados são desbalanceados
  - ▶ Se quiser uma métrica enviesada para as classes majoritárias, use Micro.
  - ▶ Se quiser uma métrica enviesada para as classes minoritárias, use Macro.
  - ▶ Se  $\text{Micro} \ll \text{Macro} \rightarrow$  veja as classes majoritárias
  - ▶ Se  $\text{Micro} \gg \text{Macro}$



# Micro x Macro

---

- ▶ Se os dados são balanceados, Micro e Macro são aproximadamente iguais
- ▶ Se dados são desbalanceados
  - ▶ Se quiser uma métrica enviesada para as classes majoritárias, use Micro.
  - ▶ Se quiser uma métrica enviesada para as classes minoritárias, use Macro.
  - ▶ Se  $\text{Micro} \ll \text{Macro}$   $\rightarrow$  veja as classes majoritárias
  - ▶ Se  $\text{Micro} \gg \text{Macro}$   $\rightarrow$  veja as classes minoritárias



# Exercício

---

- ▶ Carregar os dados do dataset iris
- ▶ Utilizar alguns métodos de classificação e obter as métricas apresentadas



# Métricas para regressão

# Métricas para Regressão

---

- ▶ Métricas bastante semelhantes
- ▶  $R^2$  é bastante utilizado
- ▶ Outras métricas
  - ▶ Erro absoluto médio
  - ▶ Erro quadrático médio
  - ▶ Mediana do erro absoluto





# Regressões Dummy

---

- ▶ **Teste de sanidade**
  - ▶ Média (mean)
  - ▶ Mediana (median)
  - ▶ Constante (constant)
  - ▶ Quantil (quantile)



# Exercício

---

- ▶ Carregar os dados do dataset diabetes
- ▶ Utilizar alguns métodos de regressão e compará-los com os métodos Dummy



# Seleção de Modelos

# Seleção de modelos

---

- ▶ Treinar e testar nos mesmo dados
  - ▶ Overfitting
- ▶ Divisão treino\teste
  - ▶ Simples
  - ▶ Sem informação de variância da métrica
- ▶ Validação cruzada
  - ▶ Mais confiável
  - ▶ Variância da métrica
  - ▶ Pode ser combinado com estratégia de grid-search



# Seleção de modelos

---

- ▶ CV e grid search para seleção de modelos usando o conjunto de treinamento (não usar o teste)
- ▶ Na prática
  - ▶ Treinamento (ajuste dos parâmetros)
  - ▶ Validação (seleção do modelo)
  - ▶ Teste (avaliação do modelo)



# Exercício

---

- ▶ Escolha um conjunto de dados de classificação
- ▶ Use o SVM com kernel RBF
- ▶ Faça 10 divisões de treino e teste
- ▶ Para cada divisão use o CV e o grid search para encontrar o melhor conjunto de hyperparâmetros ( $C$  e  $\gamma$ )
- ▶ Teste o modelo encontrado no conjunto de teste
- ▶ Repita o processo para cada um das 10 divisões.  
Apresente o resultado em termos de todas as métricas de classificação vistas hoje

