

# Machine Learning

## Dia 5 - Métricas e Seleção de Modelos

---

ImageU - Grupo de Pesquisa em Machine Learning e Visão Computacional

<https://imageu.github.io/>

Curso de Verão 2022

Instituto de Matemática e Estatística - IME USP



1. Seleção de Modelos
2. Como escolher o melhor modelo?
3. Regularização
4. Métricas

# Seleção de Modelos

---

- Bons modelos devem equilibrar duas forças:
  - Capacidade de explicação dos dados conhecidos
  - Capacidade de generalização para novos dados
- Para aferir estas propriedades, utilizamos:
  - Métricas
  - Diferentes “visões” do dataset

# Workflow Típico de ML

- Obtenção dos dados
- Divisão dos dados em dois conjuntos: treinamento e teste
- Treinamento de diferentes modelos com o conjunto de treinamento
- Escolha do melhor modelo
- Avaliação do melhor modelo no conjunto de teste

# Workflow Típico de ML

- Obtenção dos dados
- Divisão dos dados em dois conjuntos: treinamento e teste
- Treinamento de diferentes modelos com o conjunto de treinamento
- Escolha do melhor modelo
  - Como escolher o melhor modelo?
- Avaliação do melhor modelo no conjunto de teste

- Antes, uma palavrinha sobre viés (bias)
  - Confirmation Bias: quando selecionamos os dados e/ou resultados que mais favorecem os nossos experimentos;
  - Selection Bias: problema na seleção dos dados que prejudica a obtenção de amostras representativas;
  - Survival Bias: quando nos concentramos em apenas uma parte do dataset, ignorando que há outros dados

- Antes, uma palavrinha sobre viés (bias)
  - Confirmation Bias: *tenho uma opinião, e só busco informações que a confirmam;*
  - Selection Bias: *pesquisa eleitoral realizada somente em regiões que sabidamente favorecem um candidato;*
  - Survival Bias: *todo mundo que morreu de câncer bebeu água, portanto água deve causar câncer. (e quem não morreu de câncer?)*
    - Serve para lembrar que *correlação  $\neq$  causalidade :-)*



# Como escolher o melhor modelo?

---

# Definições

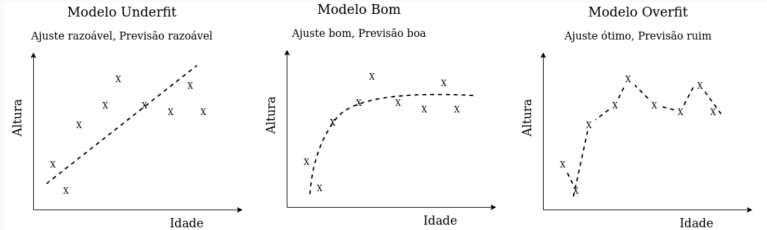
- Seja  $E_{in}$  o erro de um modelo sobre o conjunto de treinamento (*in-sample*)
- Seja  $E_{out}$  o erro de um modelo sobre o conjunto de teste (*out-of-sample*)
- Baixo  $E_{in}$  indica que o modelo possui boa capacidade de explicação dos dados de treinamento
- Baixo  $E_{out}$  indica que o modelo possui boa capacidade de generalização para os dados de teste (e com sorte, para outros dados não vistos durante o treinamento)
- O modelo ideal é aquele em que  $E_{in} \approx E_{out}$
- E quando isso não ocorre?

# Overfitting e Underfitting

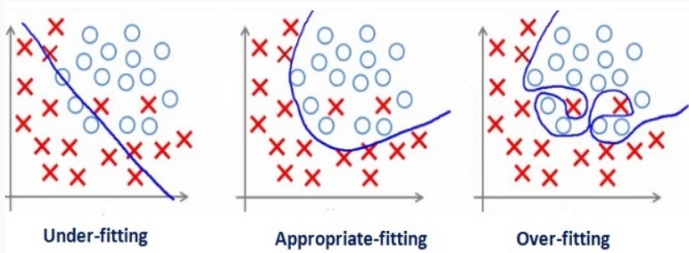
- $E_{in} \ll E_{out}$  é sinal de overfitting
- Overfitting é quando o modelo está sobreajustado aos dados de treinamento
- Dizemos popularmente que o modelo “memorizou” os dados de treinamento, e portanto não tem boa capacidade de generalização
- $E_{in}$  alto é sinal de underfitting
- Underfitting ocorre quando:
  - O modelo é muito simples para os dados (mais raro)
  - Os dados possuem algum problema, como atributos não informativos, erros de rotulação, etc (mais comum)

# Overfitting e Underfitting

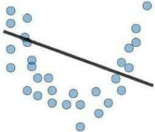


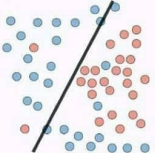
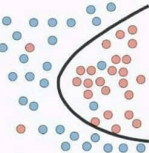
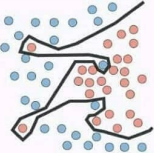
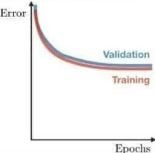
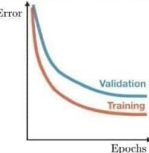
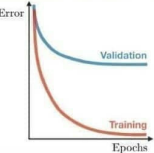
- Em modelos de Regressão:



- Em problemas de Classificação:



# Overfitting e Underfitting

	Underfitting	Just right	Overfitting
Symptoms	<ul style="list-style-type: none"><li>• High training error</li><li>• Training error close to test error</li><li>• High bias</li></ul>	<ul style="list-style-type: none"><li>• Training error slightly lower than test error</li></ul>	<ul style="list-style-type: none"><li>• Very low training error</li><li>• Training error much lower than test error</li><li>• High variance</li></ul>
Regression illustration			
Classification illustration			
Deep learning illustration			

# Overfitting e Underfitting

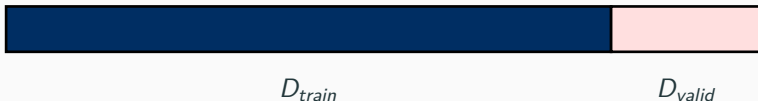
- O que fazer em casos de Overfitting? (mais comuns)
  - Utilizar um modelo mais simples
  - Treinar por menos iterações
  - Aplicar regularização\*
  - Validação cruzada (cross validation)\*
  - Conseguir mais dados\*
- O que fazer em casos de Underfitting? (mais raros)
  - Checar qualidade dos dados
  - Utilizar modelos mais complexos
  - Conseguir mais dados
- Note que conseguir mais dados quase sempre ajuda em problemas de ML

# Cross Validation (CV)

- Técnica para estimar  $E_{out}$  com parte dos dados de treinamento
- Ajuda a evitar o overfitting
- Estratégias mais comuns:
  - Hold-out (utilizada em Deep Learning)
  - k-Fold Cross Validation (utilizada em todo o resto)

# Hold-out

- Reservamos uma parte do conjunto de treinamento para validação
- Estimamos  $E_{out}$  utilizando o conjunto de validação. Chamamos essa estimativa  $E_{val}$



- Vantagens:
  - Requer treinar apenas um modelo para cada conjunto de parâmetros
- Desvantagens:
  - Se  $D_{valid}$  for muito pequeno,  $E_{val}$  não será uma boa estimativa
  - Se  $D_{valid}$  for muito grande, sobrarão poucos dados para treinamento
  - $E_{val}$  pode ser uma estimativa otimista de  $E_{out}$ , dependendo da escolha de  $D_{valid}$



# k-Fold Cross Validation

- Repetimos  $k$  vezes a estratégia usada em Hold-out:

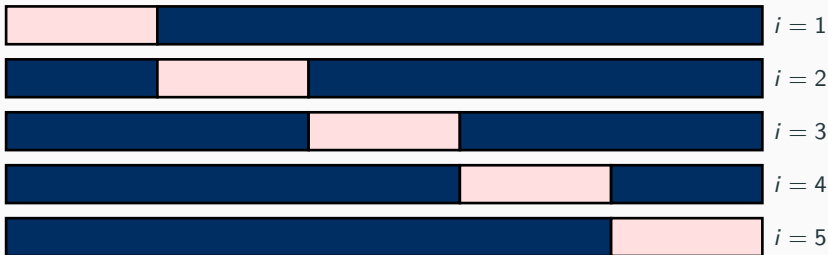
- Dividimos  $D_{train}$  em  $k$  partes iguais  $D_1, D_2, \dots, D_k$

- Repetimos o treinamento  $k$  vezes

- Na rodada  $i$ ,  $D_{train}^{(i)} = D \setminus D_i$  e  $D_{val}^{(i)} = D_i$

- $$E_{cv} = \frac{1}{k} \sum_{i=1}^k E_{val}^{(i)}$$

- Exemplo com  $k = 5$ :



- Vantagem:  $E_{cv}$  é uma estimativa mais robusta de  $E_{out}$  do que  $E_{val}$
- Desvantagem: É preciso treinar  $k$  modelos

# Workflow Típico de ML - Ajustado

- Obtenção dos dados
- Divisão dos dados em dois conjuntos: treinamento e teste
- Treinamento de diferentes modelos com o conjunto de treinamento
  - Utilizando Cross Validation
- Escolha do melhor modelo
- Avaliação do melhor modelo no conjunto de teste
  - Depois de re-treinar o melhor modelo com todo o conjunto de treinamento

# Regularização

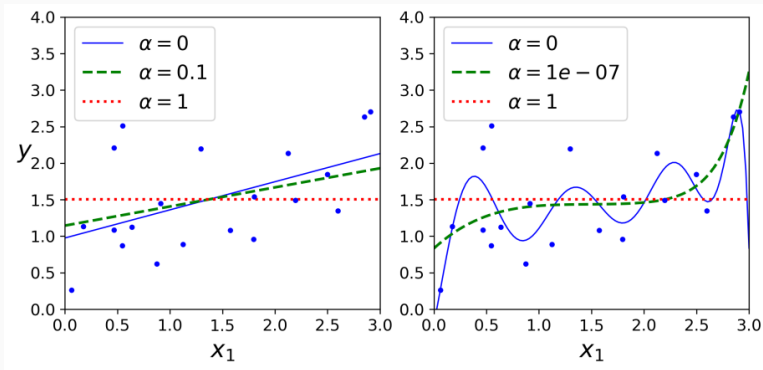
---

# Regularização $L1$ e $L2$

- Conjunto de técnicas que ajudam o algoritmo de aprendizagem a criar modelos menos complexos, e portanto menos propensos a overfitting
- Duas técnicas tradicionais, regularização  $L1$  e  $L2$ , que adicionam termos extras à função de custo, cuja intensidade é controlada por  $\alpha$ :
  - $L1$ :  $J(w) + \alpha \sum_{w \in \mathbf{w}} |w|$
  - $L2$ :  $J(w) + \alpha \sum_{w \in \mathbf{w}} w^2$
- As duas técnicas tornam o problema de otimização mais difícil, mas afetam o modelo de formas diferentes:
  - $L1$  faz com que os valores de  $\mathbf{w}$  menos importantes sejam zerados (*weight pruning*)
  - $L2$  faz com que os valores de  $\mathbf{w}$  sejam reduzidos tal que todos tenham uma contribuição mínima (*weight decay*)
- Na prática:
  - $L1$  gera modelos mais fáceis de interpretar, por possuírem menos pesos
  - $L2$  gera modelos com desempenho melhor

# Regularização $L1$ e $L2$

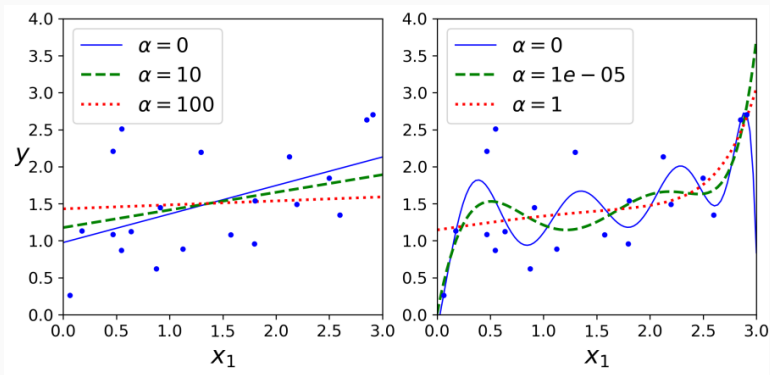
- Exemplos de regularização  $L1$  (modelo linear e polinômio de grau 10):



Fonte: *Hands-On Machine Learning with Scikit-Learn and Tensorflow*

# Regularização $L1$ e $L2$

- Exemplos de regularização  $L2$  (modelo linear e polinômio de grau 10):



Fonte: *Hands-On Machine Learning with Scikit-Learn and Tensorflow*

# Outras técnicas que possuem efeito regularizador

- Conseguir mais dados (nem sempre possível)
- Criar mais dados: Data Augmentation!
  - Criação de amostras sintéticas mas verossímeis a partir das amostras originais:
  - No caso de dados tabulares, criar versões dos dados existentes com ruído
  - No caso de imagens, criar versões rotacionadas, com alterações de brilho, contraste, etc

# Métricas

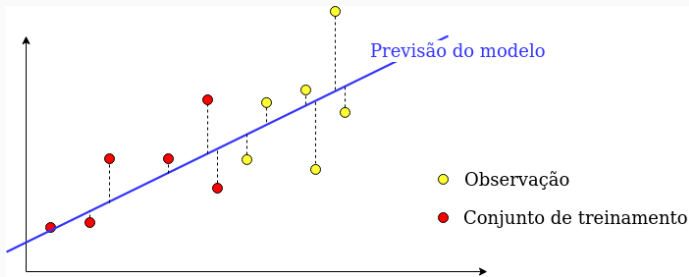
---



# Métricas - Regressão

- Erro (diferença) entre o valor observado e o valor previsto

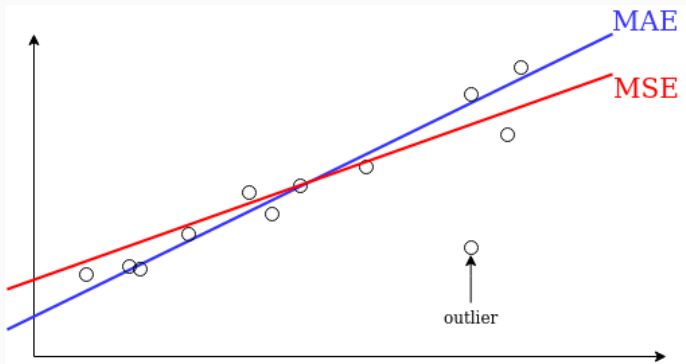
$$E_i = y_i - \hat{y}_i$$



- Erro pode ser agregado de diversas formas (MSE, MAE, Huber, logcosh, etc)
- MSE e MAE são as mais comuns

# Métricas - Regressão

- Mean Square Error - MSE:  $\frac{1}{N} \sum_{i=1}^n (y_i - \hat{y}_i)^2$
- Mean Absolute Error - MAE:  $\frac{1}{N} \sum_{i=1}^n |y_i - \hat{y}_i|$



- Principal diferença: MAE é menos sensível a outliers

# Métricas - Classificação Binária

- Duas classes possíveis: Positivo/Negativo
- Quatro possíveis diagnósticos:
  - Verdadeiro Positivo (TP)
  - Falso Positivo (FP)
  - Falso Negativo (FN)
  - Verdadeiro Negativo (TN)

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

# Métricas - Classificação Binária

## Matriz de confusão

	True condition			
	Condition positive	Condition negative		
Total population			Prevalence = $\frac{\Sigma \text{Condition positive}}{\Sigma \text{Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{True positive} + \Sigma \text{True negative}}{\Sigma \text{Total population}}$
Predicted condition positive	<b>True positive</b> , Power	<b>False positive</b> , Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{True positive}}{\Sigma \text{Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{False positive}}{\Sigma \text{Predicted condition positive}}$
Predicted condition negative	<b>False negative</b> , Type II error	<b>True negative</b>	False omission rate (FOR) = $\frac{\Sigma \text{False negative}}{\Sigma \text{Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{True negative}}{\Sigma \text{Predicted condition negative}}$
	True positive rate (TPR), Recall, Sensitivity, probability of detection = $\frac{\Sigma \text{True positive}}{\Sigma \text{Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{False positive}}{\Sigma \text{Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	<b>Diagnostic odds ratio (DOR)</b> = $\frac{\text{LR+}}{\text{LR-}}$  <b>F<sub>1</sub> score =</b> $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
	False negative rate (FNR), Miss rate = $\frac{\Sigma \text{False negative}}{\Sigma \text{Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) = $\frac{\Sigma \text{True negative}}{\Sigma \text{Condition negative}}$	Negative likelihood ratio (LR-) = $\frac{\text{FNR}}{\text{TNR}}$	

# Métricas - Classificação Binária

$$\text{Recall} = \frac{TP}{TP+FN}$$

	True condition			
	Condition positive	Condition negative		
Total population			Prevalence = $\frac{\Sigma \text{Condition positive}}{\Sigma \text{Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{True positive} + \Sigma \text{True negative}}{\Sigma \text{Total population}}$
Predicted condition positive	<b>True positive</b> , Power	<b>False positive</b> , Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{True positive}}{\Sigma \text{Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{False positive}}{\Sigma \text{Predicted condition positive}}$
Predicted condition negative	<b>False negative</b> , Type II error	<b>True negative</b>	False omission rate (FOR) = $\frac{\Sigma \text{False negative}}{\Sigma \text{Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{True negative}}{\Sigma \text{Predicted condition negative}}$
	True positive rate (TPR), Recall, Sensitivity, probability of detection = $\frac{\Sigma \text{True positive}}{\Sigma \text{Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{False positive}}{\Sigma \text{Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{TPR}{FPR}$	Diagnostic odds ratio (DOR) = $\frac{LR+}{LR-}$  F1 score = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
	False negative rate (FNR), Miss rate = $\frac{\Sigma \text{False negative}}{\Sigma \text{Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) = $\frac{\Sigma \text{True negative}}{\Sigma \text{Condition negative}}$	Negative likelihood ratio (LR-) = $\frac{FNR}{TNR}$	

# Métricas - Classificação Binária

$$\text{False Positive Rate} = \frac{FP}{FP + TN}$$

Total population	True condition		Prevalence = $\frac{\Sigma \text{Condition positive}}{\Sigma \text{Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{True positive} + \Sigma \text{True negative}}{\Sigma \text{Total population}}$	
	Condition positive	Condition negative			
Predicted condition positive	<b>True positive</b> , Power	<b>False positive</b> , Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{True positive}}{\Sigma \text{Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{False positive}}{\Sigma \text{Predicted condition positive}}$	
Predicted condition negative	<b>False negative</b> , Type II error	<b>True negative</b>	False omission rate (FOR) = $\frac{\Sigma \text{False negative}}{\Sigma \text{Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{True negative}}{\Sigma \text{Predicted condition negative}}$	
	True positive rate (TPR), Recall, Sensitivity, probability of detection = $\frac{\Sigma \text{True positive}}{\Sigma \text{Condition positive}}$	<b>False positive rate (FPR)</b> , Fall-out, probability of false alarm = $\frac{\Sigma \text{False positive}}{\Sigma \text{Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{TPR}{FPR}$	Diagnostic odds ratio (DOR) = $\frac{LR+}{LR-}$	F <sub>1</sub> score = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
	False negative rate (FNR), Miss rate = $\frac{\Sigma \text{False negative}}{\Sigma \text{Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) = $\frac{\Sigma \text{True negative}}{\Sigma \text{Condition negative}}$	Negative likelihood ratio (LR-) = $\frac{FNR}{TNR}$		

# Métricas - Classificação Binária

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

Total population	True condition		Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
	Condition positive	Condition negative			
Predicted condition positive	<b>True positive</b> , Power	<b>False positive</b> , Type I error	<b>Positive predictive value (PPV),</b> Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	<b>False discovery rate (FDR) =</b> $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
Predicted condition negative	<b>False negative</b> , Type II error	<b>True negative</b>	<b>False omission rate (FOR) =</b> $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	<b>Negative predictive value (NPV) =</b> $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
	<b>True positive rate (TPR),</b> Recall, Sensitivity, probability of detection = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	<b>False positive rate (FPR),</b> Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	<b>Positive likelihood ratio (LR+)</b> = $\frac{TPR}{FPR}$	<b>Diagnostic odds ratio (DOR)</b> = $\frac{LR+}{LR-}$	<b>F1 score =</b> $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
	<b>False negative rate (FNR),</b> Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	<b>Specificity (SPC),</b> Selectivity, <b>True negative rate (TNR)</b> = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	<b>Negative likelihood ratio (LR-)</b> = $\frac{FNR}{TNR}$		

# Métricas - Classificação Binária

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}$$

Total population	True condition		Prevalence = $\frac{\Sigma \text{Condition positive}}{\Sigma \text{Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{True positive} + \Sigma \text{True negative}}{\Sigma \text{Total population}}$	
	Condition positive	Condition negative			
Predicted condition positive	<b>True positive</b> , Power	<b>False positive</b> , Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{True positive}}{\Sigma \text{Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{False positive}}{\Sigma \text{Predicted condition positive}}$	
Predicted condition negative	<b>False negative</b> , Type II error	<b>True negative</b>	False omission rate (FOR) = $\frac{\Sigma \text{False negative}}{\Sigma \text{Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{True negative}}{\Sigma \text{Predicted condition negative}}$	
	True positive rate (TPR), Recall, Sensitivity, probability of detection = $\frac{\Sigma \text{True positive}}{\Sigma \text{Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{False positive}}{\Sigma \text{Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR-}}$	F <sub>1</sub> score = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
	False negative rate (FNR), Miss rate = $\frac{\Sigma \text{False negative}}{\Sigma \text{Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) = $\frac{\Sigma \text{True negative}}{\Sigma \text{Condition negative}}$	Negative likelihood ratio (LR-) = $\frac{\text{FNR}}{\text{TNR}}$		



# Métricas - Classificação Binária

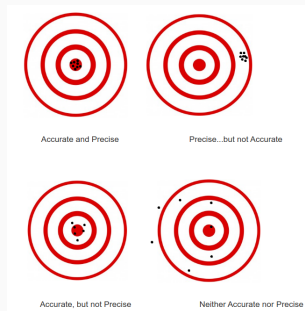
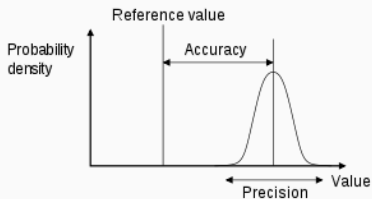
$$F1\text{-score} = 2 \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

	True condition			
	Condition positive	Condition negative		
Total population			Prevalence = $\frac{\Sigma \text{Condition positive}}{\Sigma \text{Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{True positive} + \Sigma \text{True negative}}{\Sigma \text{Total population}}$
Predicted condition positive	<b>True positive</b> , Power	<b>False positive</b> , Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{True positive}}{\Sigma \text{Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{False positive}}{\Sigma \text{Predicted condition positive}}$
Predicted condition negative	<b>False negative</b> , Type II error	<b>True negative</b>	False omission rate (FOR) = $\frac{\Sigma \text{False negative}}{\Sigma \text{Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{True negative}}{\Sigma \text{Predicted condition negative}}$
	True positive rate (TPR), Recall, Sensitivity, probability of detection = $\frac{\Sigma \text{True positive}}{\Sigma \text{Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{False positive}}{\Sigma \text{Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR-}}$
	False negative rate (FNR), Miss rate = $\frac{\Sigma \text{False negative}}{\Sigma \text{Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) = $\frac{\Sigma \text{True negative}}{\Sigma \text{Condition negative}}$	Negative likelihood ratio (LR-) = $\frac{\text{FNR}}{\text{TNR}}$	

$$F_1 \text{ score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

# Métricas - Classificação Binária

- Acurácia vs Precisão
  - Acurácia: quanto a estimativa é próxima do real
  - Precisão: quão reprodutível é o resultado
- Pense em média e desvio padrão:



Fonte: Wikipedia,

<https://blog.minitab.com/blog/real-world-quality-improvement/accuracy-vs-precision-whats-the-difference>

# Métricas - Classificação Multiclasse

- Micro-averaging
  - Calcular  $TP_j$ ,  $FP_j$ ,  $TN_j$ ,  $FN_j$  individualmente, por classe  $j$  (uma classe contra o restante)
  - Somá-los para obter  $TP$ ,  $FP$ ,  $TN$ ,  $FN$ , e calcular a métrica
  - Atribui mesmo peso para as métricas  $\rightsquigarrow$  classes maiores dominam
- Macro-averaging
  - Calcular a métrica para cada classe e depois calcular a média
  - Atribui mesmo peso para as classes  $\rightsquigarrow$  mesma importância para todas as classes
- Há controvérsias sobre qual é melhor
- Macro-averaging é o mais comum (usado no scikit-learn e no Keras)

# Classes maiores e menores?

- Datasets com classes desbalanceadas são sempre um problema
- Adiciona dificuldades no treinamento e na avaliação
- No caso de classificação binária é mais fácil tratar; no caso multi-classes nem tanto

# Próxima Aula: Pré-processamento, Regressão Logística Multi-classe

