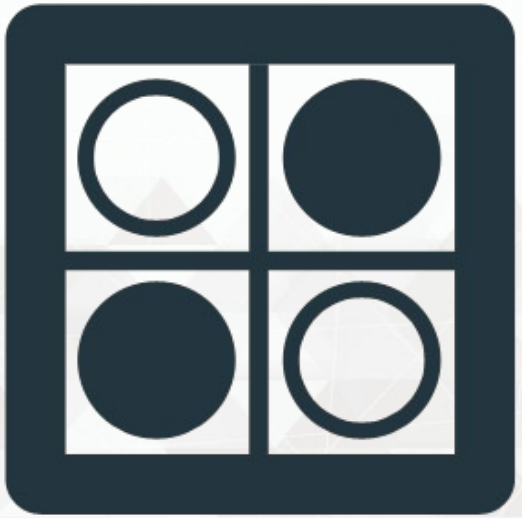
BigML
Avaliação de Modelos

2023

<lapti>

Programa

- Avaliação de modelos
  - Tipos
- Métricas
- Visualização
  - Confusion matrix
  - Curvas
- Riscos



Avaliando um modelo

Avaliações

- Comparação de desempenho de modelos de classificação e regressão
- Objetivos
  - Obter estimativa do desempenho do modelo em predição
    - Realizar predições para novas instâncias que o modelo nunca viu antes
  - Fornecer um framework para comparar modelos usando configurações e algoritmos diferentes
    - Para identificar o modelo com o melhor desempenho preditivo
- Single (training/test split) e cross-validation

Avaliações

- Single evaluation
  - Testar predições para instâncias já classificadas, mas não vistas pelo modelo
  - Training/test splits
- Cross-validation
  - Divisão dos dados classificados em vários subconjuntos
    - Usar um para avaliação e o restante para treinamento, testar todas as combinações
    - Resultado da avaliação será a média de todas
  - Basicamente uma avaliação single executada diversas vezes em diversos splits
  - K-fold cross-validation

Métricas de avaliação

- Imagine um modelo que pode prever se uma transação é fraude ou é verdadeira
  - Para cada transação, prediz se é fraude/verdadeira
- Selecione a **classe positiva**
  - A classe que você está interessado, a escolha a ser pesquisada
  - Fraude, por exemplo

Métricas de avaliação

- Classe positiva escolhida: fraude
- True Positive (TP)
  - Modelo prevê fraude e a resposta correta é fraude
- True Negative (TN)
  - Modelo prevê verdadeira e a resposta correta é verdadeira
- False Positive (FP)
  - Modelo prevê fraude, mas a resposta correta é verdadeira
- False Negative (FN)
  - Modelo prevê verdadeira, mas a resposta correta é fraude
- Confusion Matrix

Quanto custa um erro?

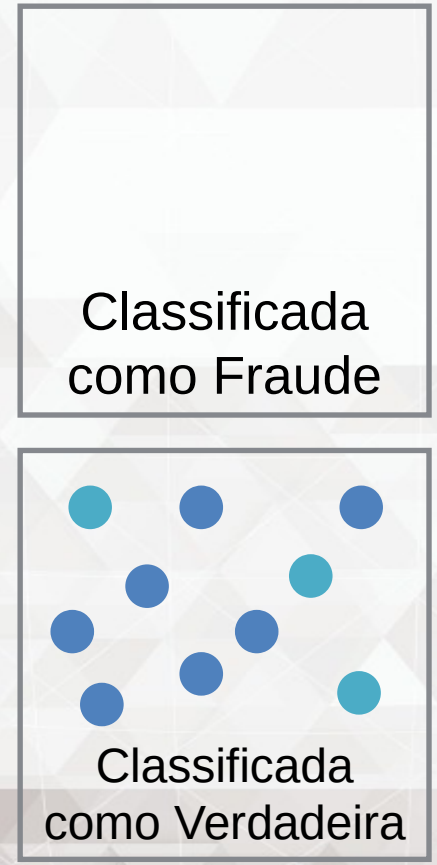
- No domínio do problema a ser resolvido pelo modelo, o que é pior, um falso positivo ou um falso negativo?
- Diagnóstico médico
  - Custo de um falso positivo
    - Paciente acaba realizando mais exames para descobrir que realmente não tem a doença
      - Baixo custo?
  - Custo de um falso negativo
    - Paciente é declarado sadio e uma doença não detectada progride e eventualmente volta a se manifestar
      - Custo alto?
- Solução
  - Selecionar um limiar para classificação positiva que tenha o trade-off apropriado em relação aos erros

Accuracy (acurácia)

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

- “Porcentagem correta” - como uma prova em sala de aula
- = 1 então sem erro nenhum
- = 0 então todos estão errados
- Intuitivo, mas nem sempre útil
- Cuidado com classes não-balanceadas
  - Ex: 90% das transações são verdadeiras e 10% são fraude
  - Um modelo simplório que SEMPRE prevê verdadeira tem **90% de acurácia**

Accuracy



● = Fraude
● = Verdadeira

TP = 0
FP = 0
TN = 7
FN = 3

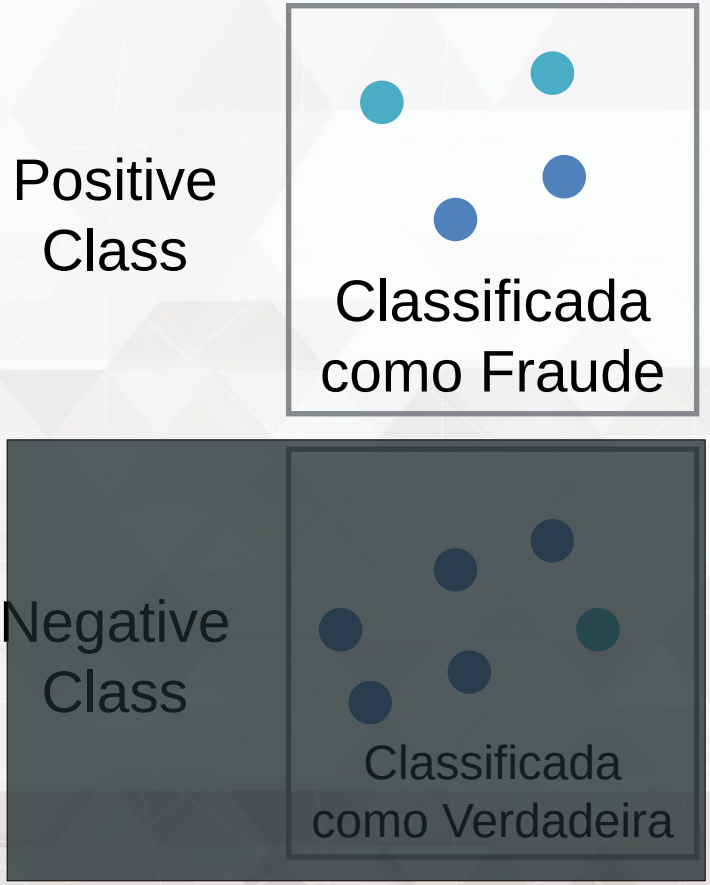
$$\frac{TP + TN}{Total} = 70\%$$

Precision (precisão)

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

- “accuracy” ou “pureza” da classe positiva
- Quão bem a classe positiva é separada da classe negativa
- Precision = 1, então sem FP
  - O modelo pode ter deixado passar algumas fraudes, mas das que foram identificados, TODAS são fraude
    - Sem erros de identificação positiva
- Precision = 0, então sem TP
  - Nenhuma das fraudes identificadas são na realidade fraudes, TODAS são erros

Precision



● = Fraude

● = Verdadeira

TP = 2
FP = 2
TN = 5
FN = 1

$$\frac{TP}{TP + FP} = 50\%$$

Recall

$$\frac{TP}{TP + FN}$$

- Sensitivity, True Positive Rate ou probabilidade de deteção
- Porcentagem da classe positiva corretamente identificada
- Medida de quão bem o modelo identifica todos os exemplos da classe positiva
- Recall = 1, então sem FN
  - TODAS as fraudes são identificadas
  - Devem existir alguns FP, então a precisão deve ser < 1
- Recall = 0, então sem TP
  - Nenhuma fraude identificada

Recall

Positive Class

Negative Class

● = Fraude

● = Verdadeira

TP = 2

FP = 2

TN = 5

FN = 1

$$\frac{TP}{TP + FN} = 66\%$$

f-Measure

$$\frac{2 * Recall * Precision}{Recall + Precision}$$

- Média harmônica de Recall e Precision
- = 1 então Recall == Precision == 1
- Se Precision OU Recall são pequenos, então f-Measure é pequena

f-Measure

Positive Class

Negative Class

= Fraude

= Verdadeira

R = 66%

P = 50%

f = 57%

Phi Coefficient

$$\frac{TP * TN - FP * FN}{\text{Sqrt}[(TP+FP)\ (TP+FN)\ (TN+FP)\ (TN+FN)]}$$

- Retorna um valor entre -1 e 1
- Se -1 então predições são oposição à realidade
- = 0 então sem correlação entre predições e a realidade
- = 1 então predições são sempre corretas

Phi Coefficient

Positive Class

Negative Class

● = Fraude
● = Verdadeira

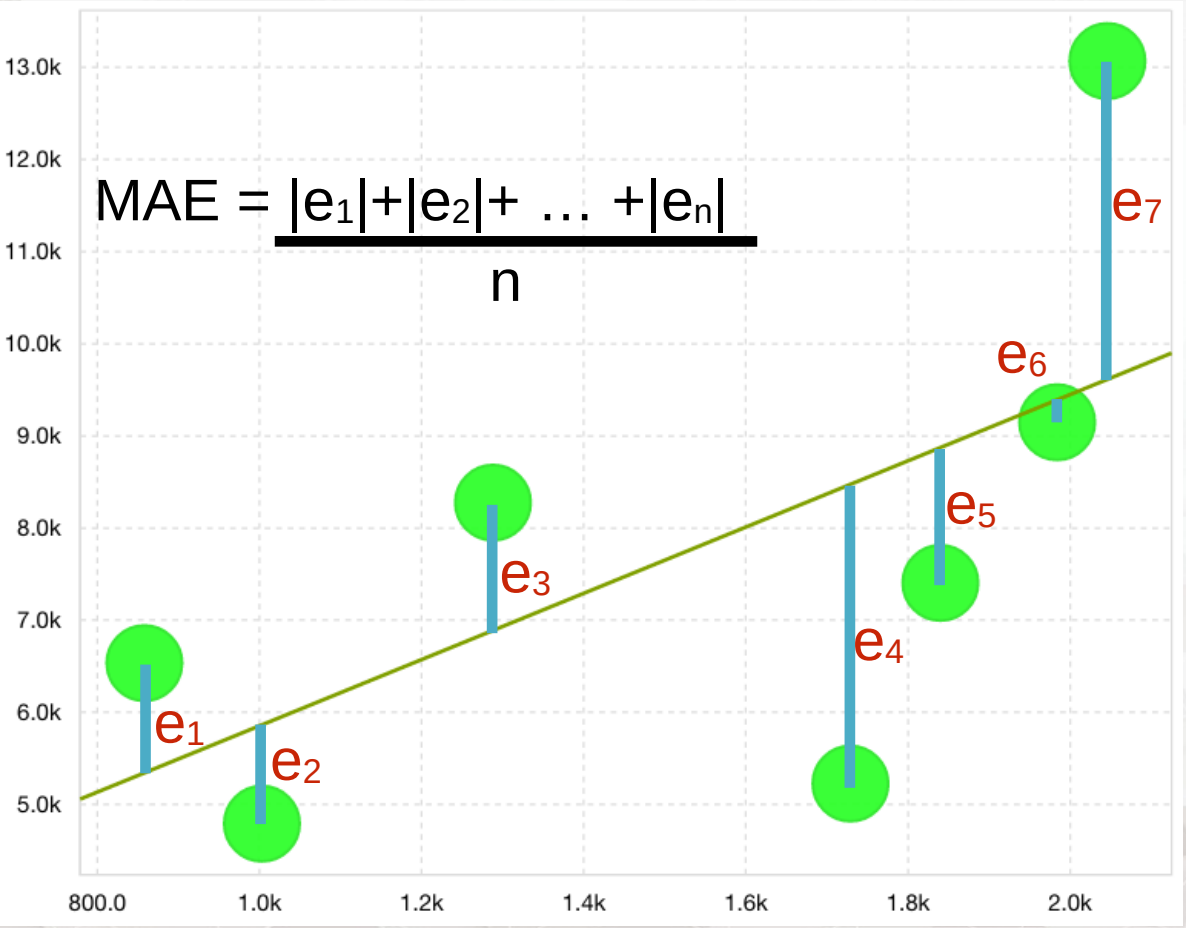
TP = 2
FP = 2
TN = 5
FN = 1

Phi = 0.356

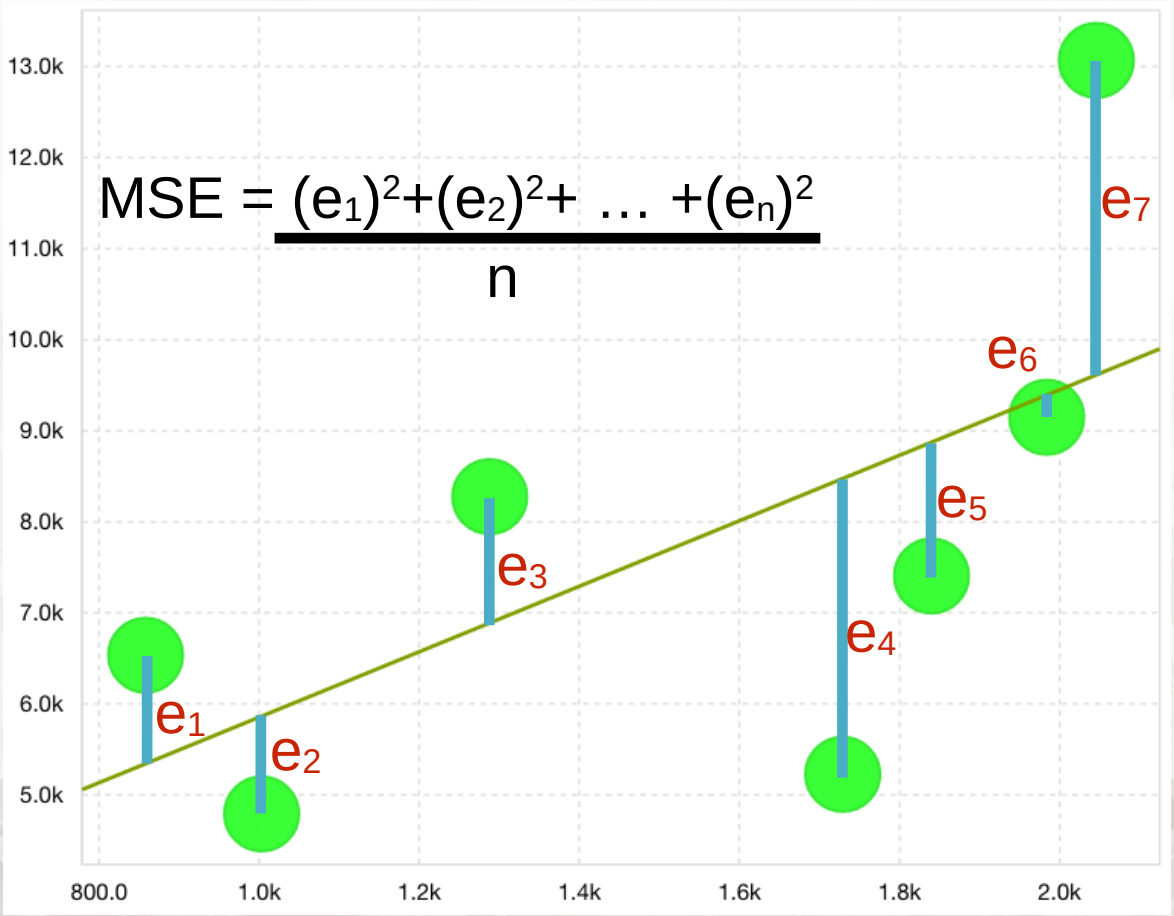
Avaliações em regressões

- Determinação de erros
  - Mean absolute error
  - Mean square error
  - R-square error

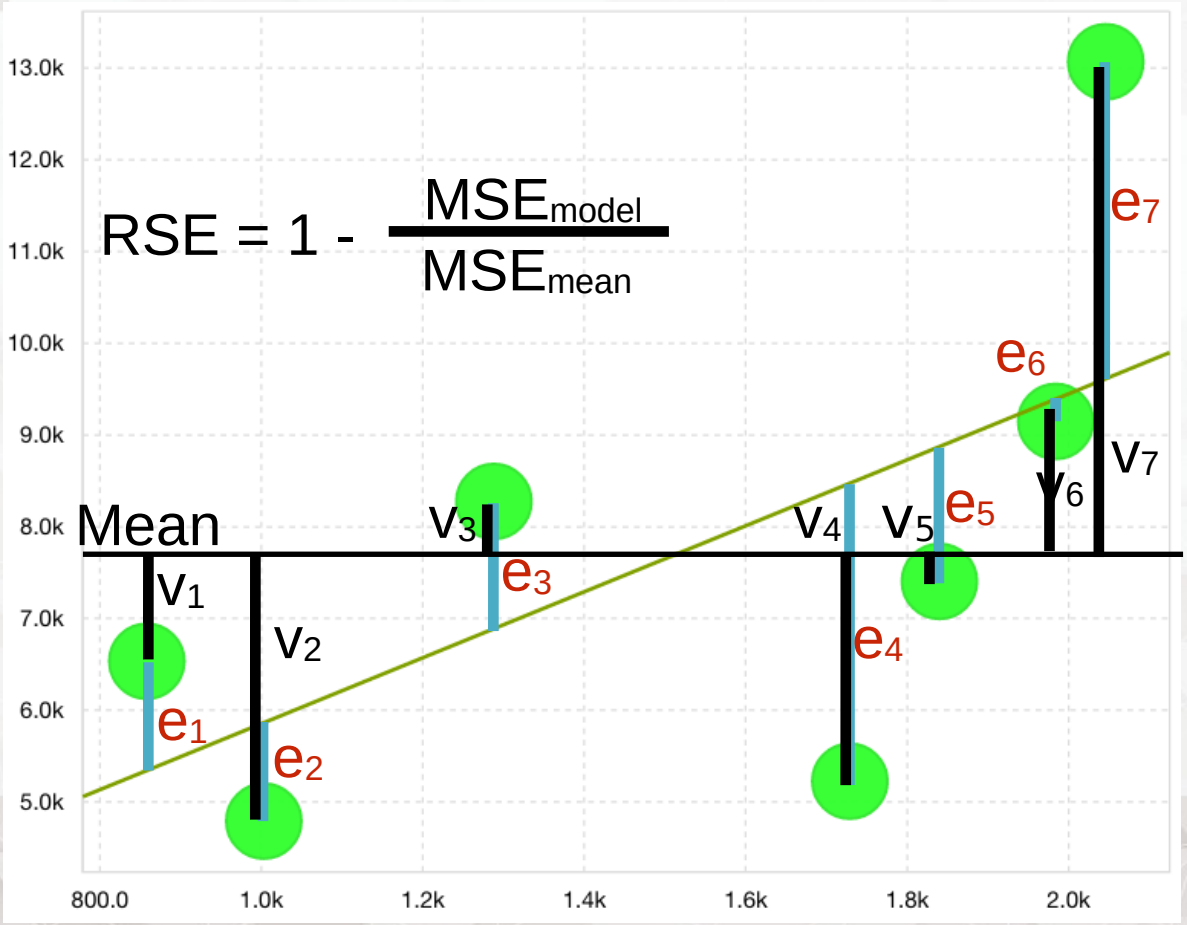
Mean Absolute Error



Mean Square Error



R-Square Error

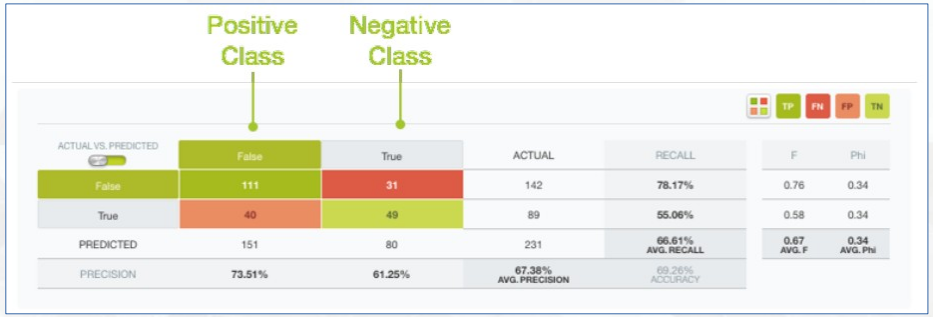


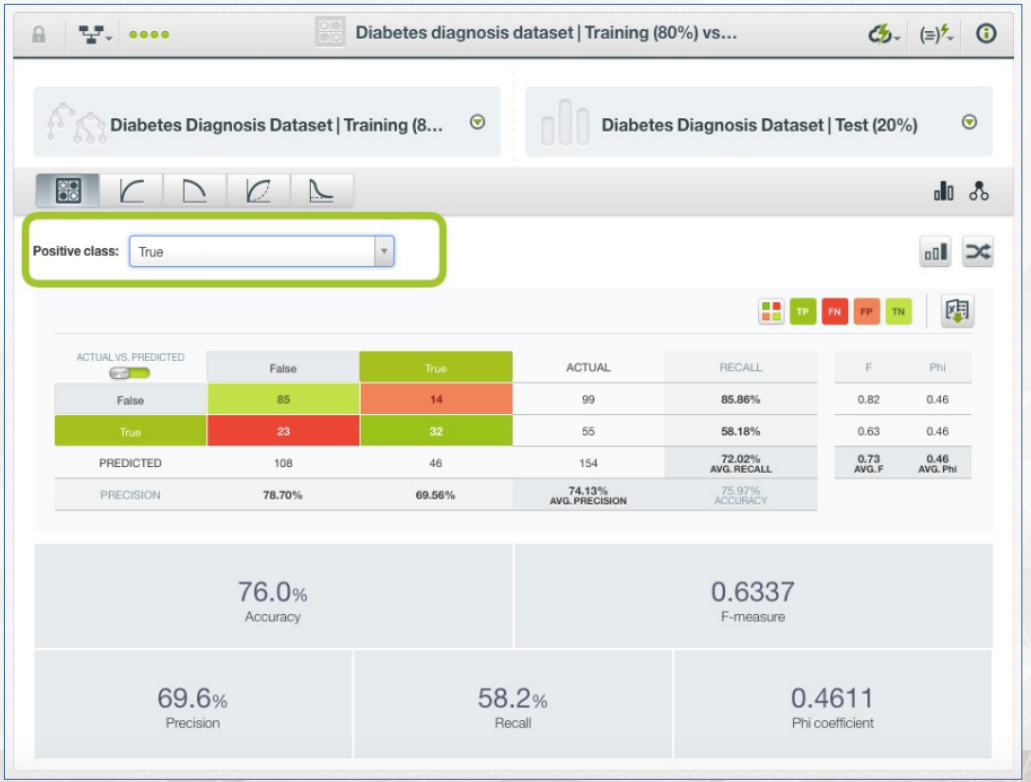
R-Squared Error

$$RSE = 1 - \frac{MSE\ model}{MSE\ mean}$$

- RSE: medida de quão melhor o modelo é do que sempre predizer com a média
- < 0 modelo é pior que a média
- = 0 modelo não é melhor que a média
- Tendendo a 1 modelo se encaixa perfeitamente nos dados

Visualização





Visualização

UTFPR
UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

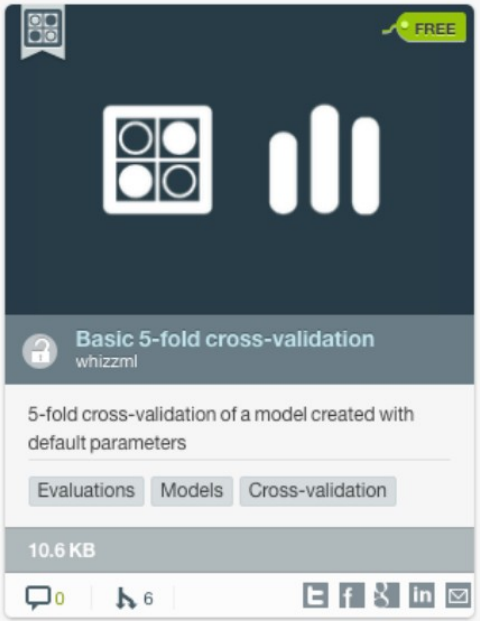
<lapti>

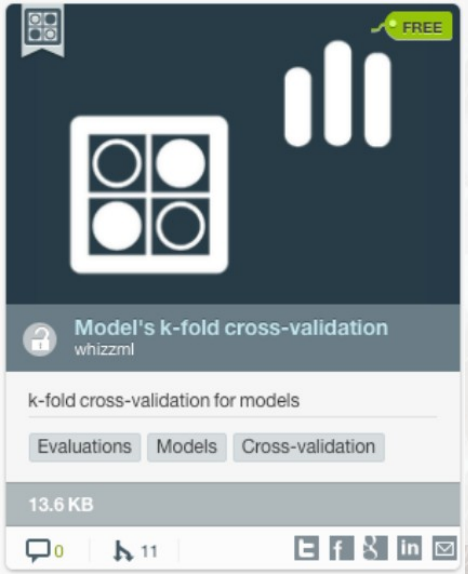
Curvas

- Precision-Recall curve
  - Trade-off entre as medidas para a classe positiva
  - Precision e recall são inversamente relacionadas
  - Quanto maior a área abaixo da curva (Area Under the Curve – AUC), melhor a predição
    - Alta precisão, alto recall
- ROC curve (Receiver Operating Characteristic)
  - Trade-off entre recall (TP rate ou sensitividade) e especificidade (TN rate)
  - Valores altos para AUC são bons, mas em casos extremos (~1) podem significar overfitting
- Gain curve
  - Relação entre porcentagem de predições corretas e porcentagem de instâncias em predição
  - Mede esforço para se chegar a um determinado acerto
- Lift curve
  - Mostra comparação do modelo com uma atribuição de classes aleatória

Cross-validation

- WhizzML script
  - BigML Gallery
  - Devem ser clonados para o dashboard para utilização
- Tipos
  - Basic 5-fold cross-validation
  - k-fold cross-validation
    - Model's, Ensemble's, Logistic regression's, Deepnet's





Riscos em avaliações

- Nunca avaliar com dados de treinamento!
  - Muitos modelos podem “memorizar” os dados de treinamento
  - Isto resulta em uma avaliação excessivamente otimista
  - Sempre usar train/test split
- Mesmo um train/test split pode não ser suficiente
  - É possível obter um test split “sortudo”
    - Aleatoriedade
  - Solução é repetir o teste várias vezes
    - Cross-validation
- Não esquecer que accuracy pode ser uma medida enganosa
  - Basicamente inútil em classes desbalanceadas

Então

- Avaliações são essenciais para validação de modelos
  - Não há como se colocar um modelo em produção sem avaliá-lo exaustivamente
  - Todo modelo em produção deve ser avaliado periodicamente
    - Com os novos dados produzidos no período
- Técnicas podem ser complexas
- Leitura dos indicadores deve ser cuidadosa
- Muito trabalho de pesquisa e desenvolvimento em avaliações de modelos de ML
  - É importante se manter atualizado

Então

- Avaliações são essenciais para validação de modelos
  - Não há como se colocar um modelo em produção sem avaliá-lo exaustivamente
  - Todo modelo em produção deve ser avaliado periodicamente
    - Com os novos dados produzidos no período
- Técnicas podem ser complexas
- Leitura dos indicadores deve ser cuidadosa
- Muito trabalho de pesquisa e desenvolvimento em avaliações de modelos de ML
  - É importante se manter atualizado

Obrigado

leandro@utfpr.edu.br
http://lapti.ct.utfpr.edu.br