BigML
Regressões Logísticas

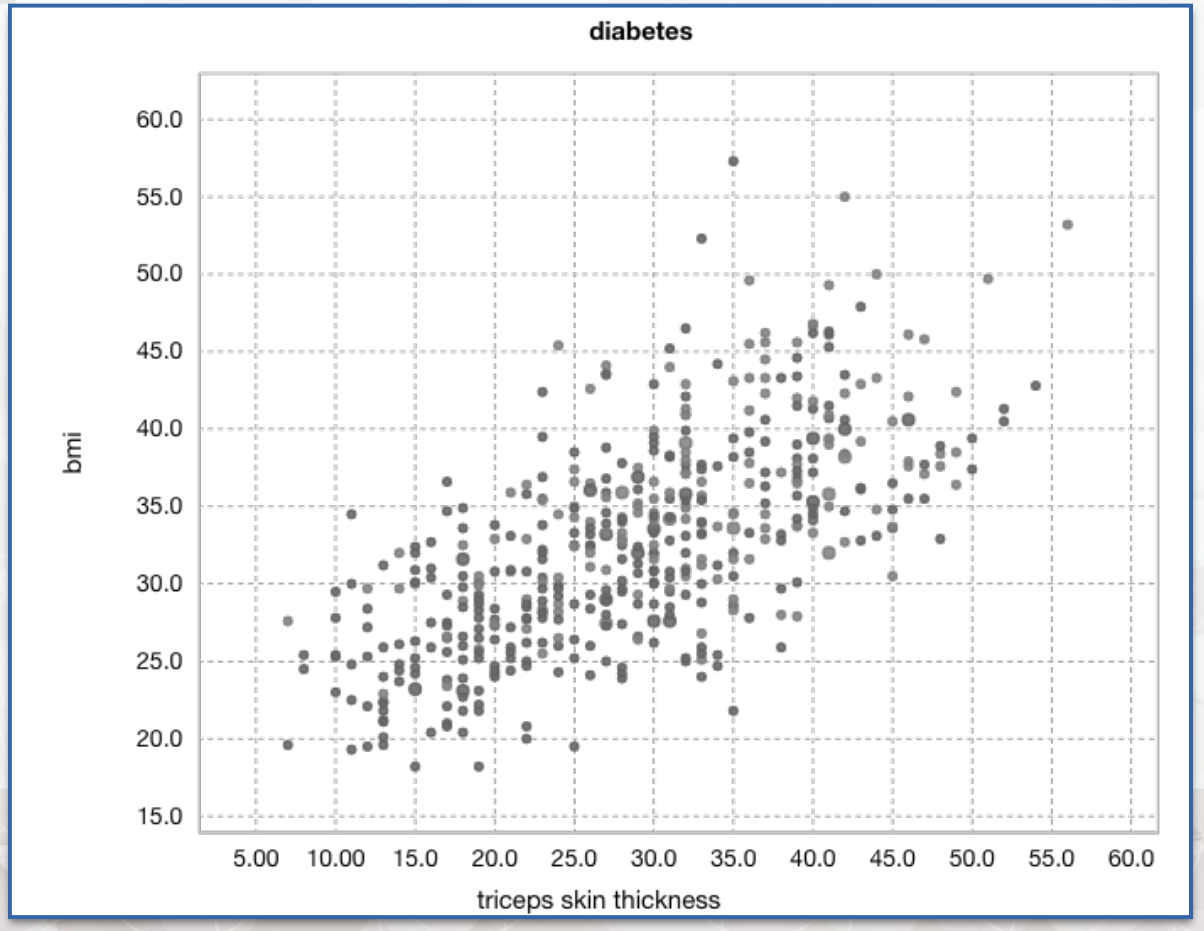
2023

<lapti>

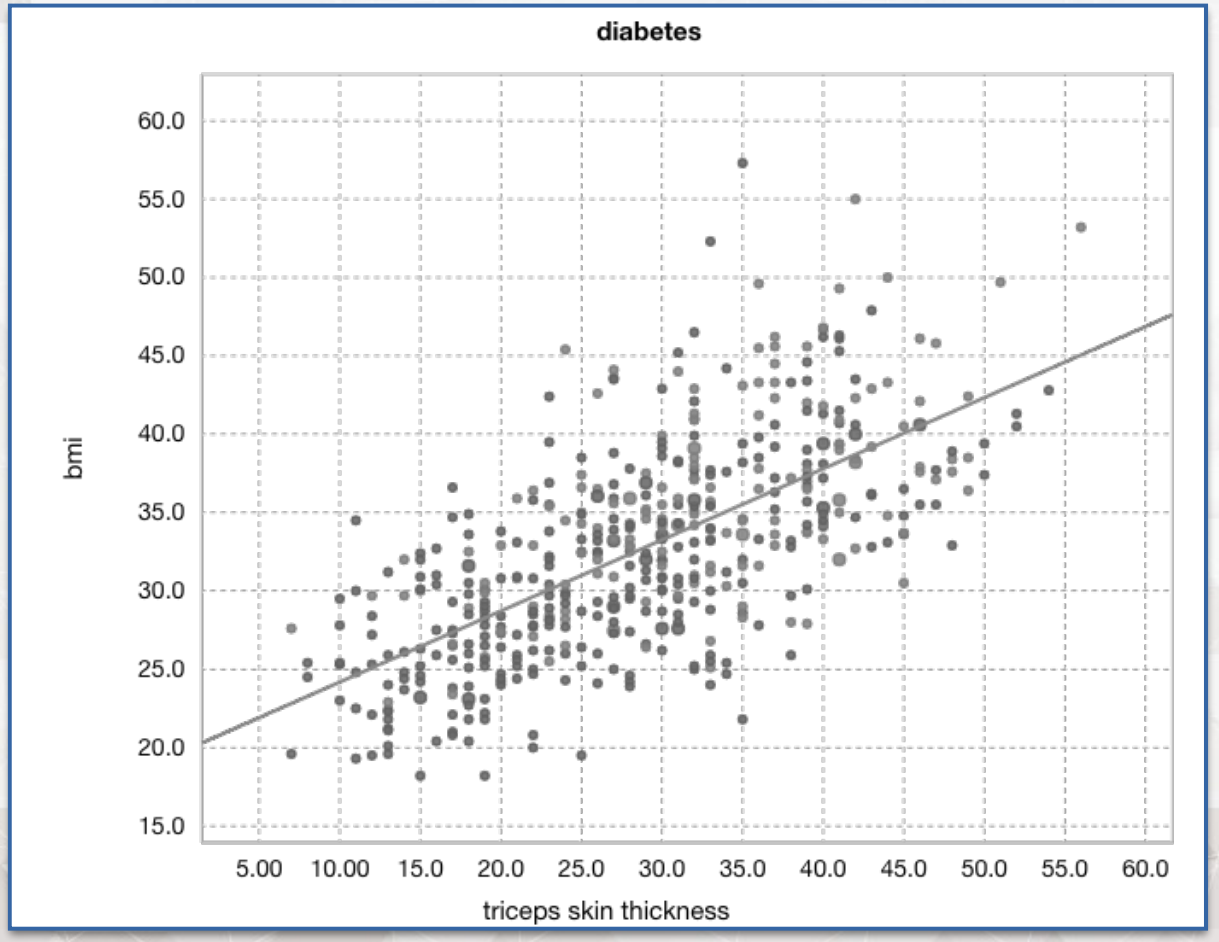
Programa

- Regressões Logísticas
  - Logistic regressions

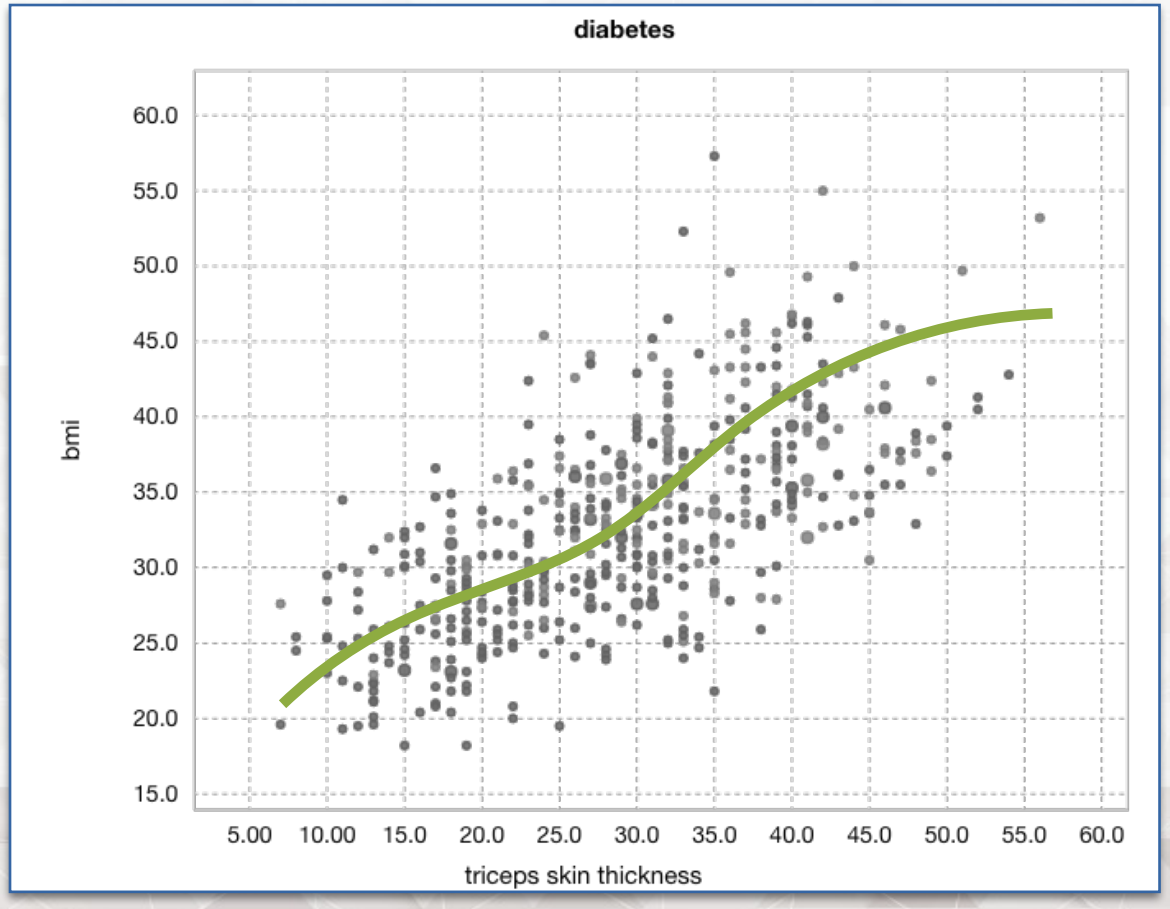
Regressão linear



Regressão linear



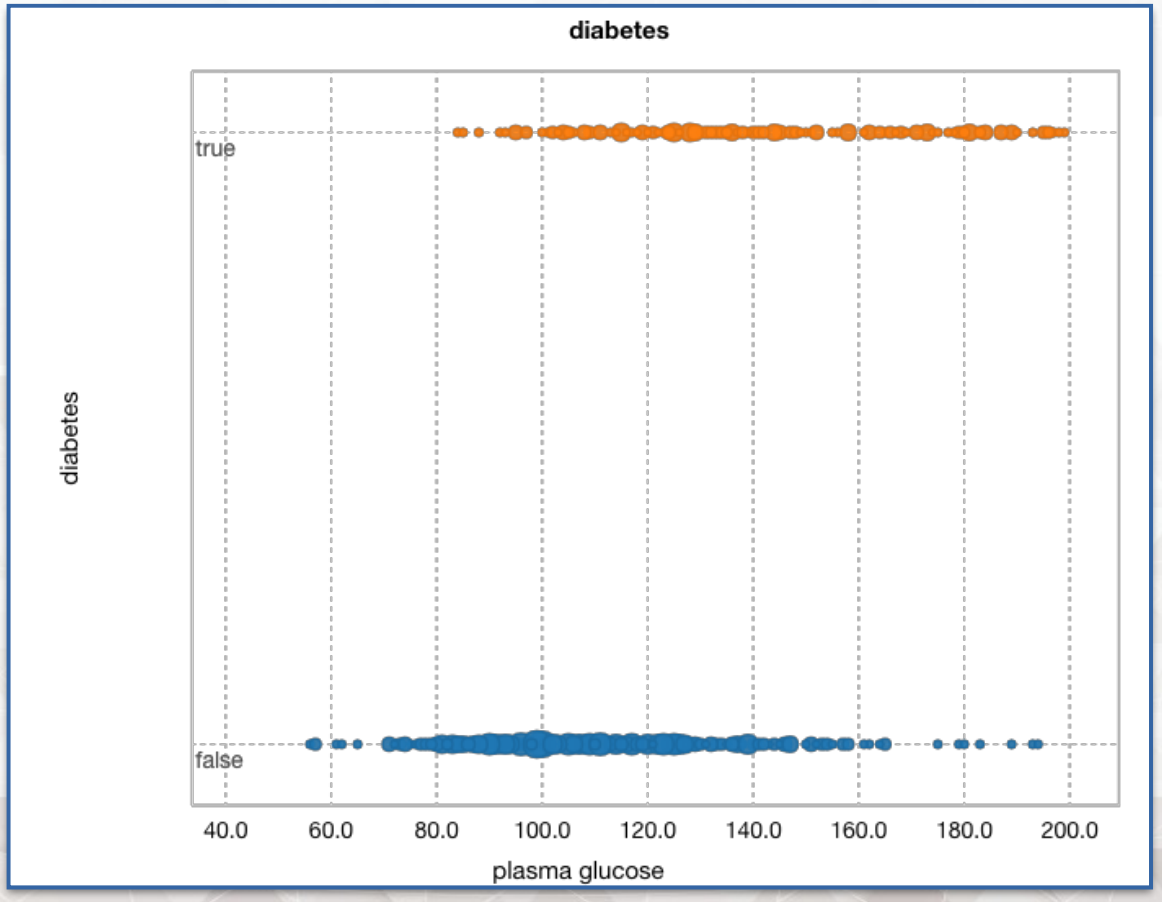
Regressão polinomial



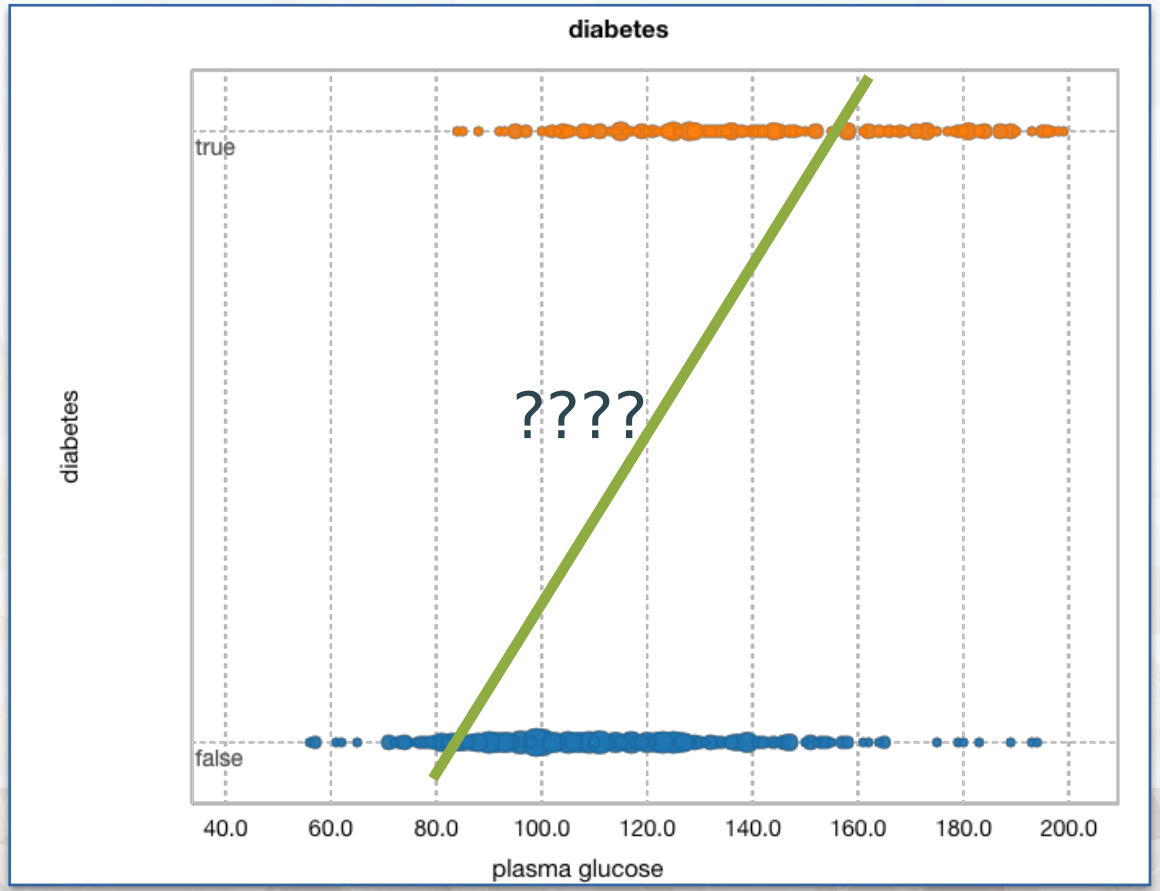
Regressões

- Modelo de classificação
  - Categorias, true/false
  - Regressões numéricas são usadas para determinação de valores
  - O termo “regressão” em LR se refere ao funcionamento interno do algoritmo, e não à saída
- Prevê a probabilidade de uma classe
  - Ex.: 75% de probabilidade de churn ser true
  - Classe com maior probabilidade é a predição
- Ajusta uma função “logit” aos dados
  - Difere de modelos que são mais flexíveis

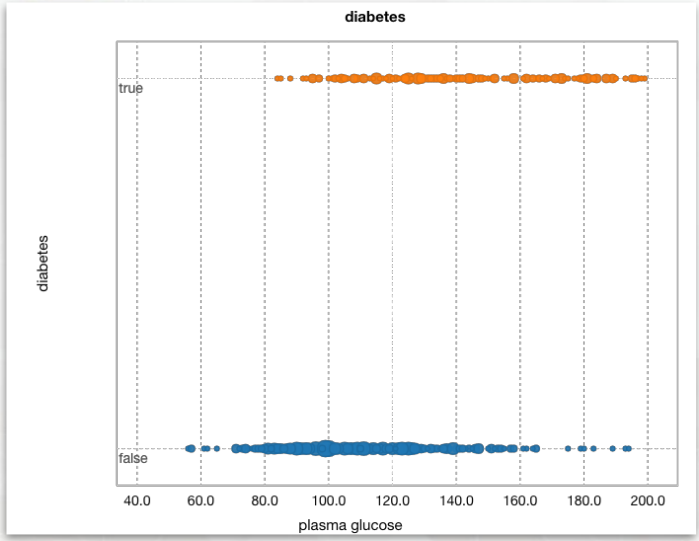
Função de dados discreta

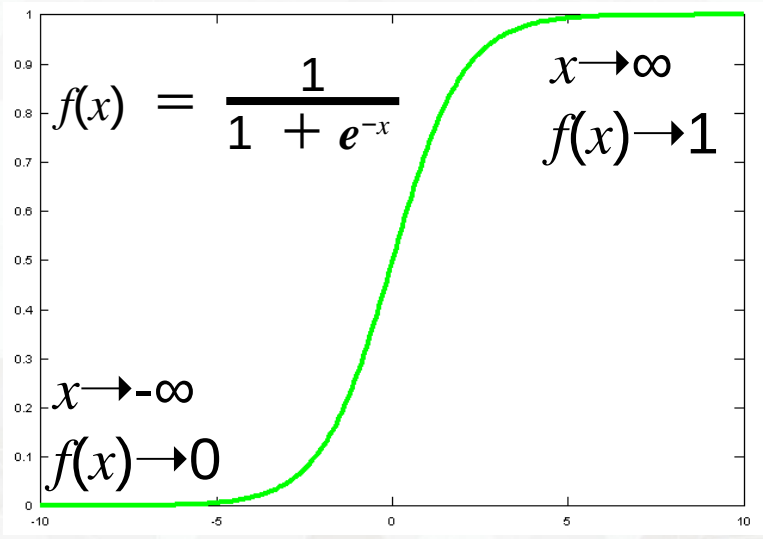


Função de dados discreta



Função logística





Goal

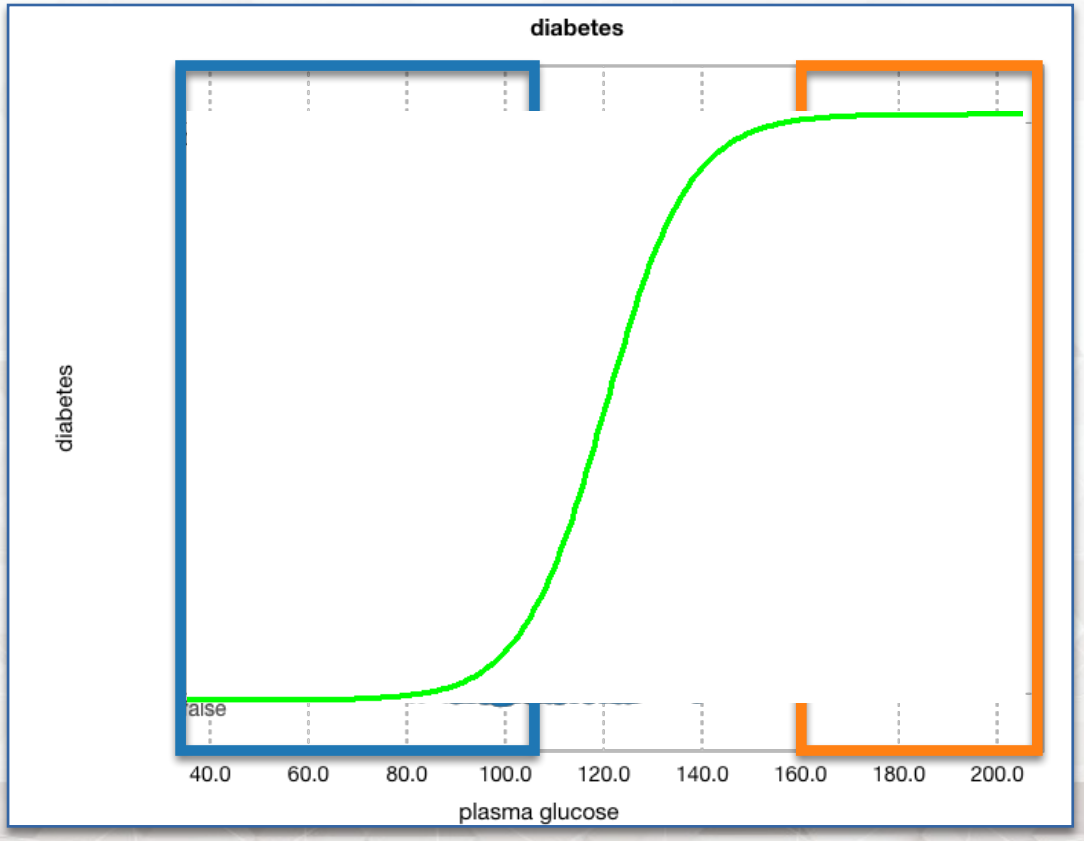
Logistic Function

- Promissor, mas ainda não “discreta”
- Muito espaço na porção intermediária

UTFPR
UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

<lapti>

Modelando probabilidades



Coeficientes LR

- LR calcula um “coeficiente” para cada feature
  - Negativo == correlação negativa
  - Positivo == correlação positiva
  - Maior == maior impacto da feature
  - Menor == menor impacto da feature
  - Não é a importância do campo
- Automaticamente inclui um coeficiente para “missing”
  - Se habilitado
- Classificação binária (true/false), coeficientes são complementares

Configuração

- Default Numeric / Missing Numeric
  - Replace / Learn / Ignore
- EPS
  - Erro mínimo entre passos para parada
- Stats
  - Estatísticas adicionais para campos
- Weights
  - Tratamento de classes desbalanceadas
- Bias
  - Permite um termo de interceptação
- Auto-scaling
  - Garante que todas as features numéricas contribuam igualmente
  - Usado quando campos diferem drasticamente em escalas
- Regularização
  - L1
    - Diminuição de coeficientes individuais
  - L2
    - Tenta aproximar todos os coeficientes a zero
      - Positivos e negativos

LR com Múltiplas Classes

- Problemas de classificação podem envolver mais de duas classes
  - Ex.: bom, regular, ruim
- Regressões Logísticas tratam múltiplas classes com one-vs-all
  - Bom / não bom
  - Regular / não regular
  - Ruim / não ruim
- Resultado final é uma combinação das regressões de cada classe individual
  - BigML faz a composição desses resultados de forma automática

Codificação de campos

- LR espera valores numéricos para realizar a regressão
- Como tratamos valores categóricos ou texto?

One-hot encoding

Classe	cor=red	cor=blue	cor=green	cor=NULL
red	1	0	0	0
blue	0	1	0	0
green	0	0	1	0
MISSING	0	0	0	1

Apenas uma feature é "hot" para cada classe
Este é o default do algoritmo

Codificação de campos

Dummy Encoding

Class	color_1	color_2	color_3
red	0	0	0
blue	1	0	0
green	0	1	0
MISSING	0	0	1

Escolhe uma *reference class*

Codificação de campos

Contrast Encoding

Class	field
red	0,5
blue	-0,25
green	-0,25
MISSING	0

"influence"
positive
negative
negative
excluded

Valores dos campos devem somar zero

Variação sem essa necessidade

Permite comparação entre as classes

Codificação de campos

- Qual usar?
  - One-hot é o default
    - Usar a não ser que exista uma necessidade específica
  - Dummy
    - Usar quando existe um grupo de controle para os dados
      - Que se torna a classe de referência
  - Contraste
    - Permite testar hipóteses específicas de relacionamentos
      - Ex.: clientes gerando um “rating” de bom / ok / ruim

Curvilinear LR

- LR espera uma relação linear entre features e objetivos
  - Na realidade, isso é bem comum
  - Relações não lineares podem impactar a qualidade do modelo
    - Quadráticas, por exemplo
- Pode ser corrigido adicionando transformações não-lineares aos dados
- Entretanto, para se saber onde aplicar as transformações é necessário:
  - Conhecimento do domínio
  - Experimentação
  - Ambos...

LR x DT

- Logistic Regression
  - Espera uma relação linear das features
  - LR traz probabilidade de uma saída discreta
  - Muitos parâmetros que podem gerar erros
  - Menor tendência a overfitting
  - **Por ajustar uma curva, trabalha melhor com menor volume de dados**
- Decision Tree
  - Se adapta bem a dados não-lineares
  - Classificação, regressão, multi-class
  - Poucos parâmetros para configuração, pouco risco
  - Maior tendência a overfitting
  - Por conta dos eixos perpendiculares, precisa de muitos dados para ajuste

LR x DT

- Logistic Regression
  - Espera uma relação linear das features
  - LR traz probabilidade de uma saída discreta
  - Muitos parâmetros que podem gerar erros
  - Menor tendência a overfitting
  - **Por ajustar uma curva, trabalha melhor com menor volume de dados**
- Decision Tree
  - Se adapta bem a dados não-lineares
  - Classificação, regressão, multi-class
  - Poucos parâmetros para configuração, pouco risco
  - Maior tendência a overfitting
  - Por conta dos eixos perpendiculares, precisa de muitos dados para ajuste

Obrigado

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