



Regressão

Notas Gerais e Comentários sobre Avaliação de Desempenho



Introdução

- A tarefa em que se prediz o valor contínuo de uma variável dá-se o nome de *estimação* (ou *regressão*) e, portanto, possui muitas características e processos em comum com a classificação
- A preparação da base de dados, a separação dos dados em treinamento e teste, a definição dos critérios de parada do algoritmo e o treinamento e teste são feitos de forma equivalente
- Uma diferença importante entre essas tarefas, entretanto, encontra-se na avaliação da saída
 - No caso dos classificadores, essa avaliação é baseada em alguma medida de acurácia do classificador, ou seja, a quantidade de objetos classificados corretamente
 - No caso dos estimadores, a qualidade é normalmente medida calculando-se uma distância ou erro entre a saída do estimador e a saída desejada



Avaliação de Desempenho

- A saída de um estimador é um valor numérico contínuo que deve ser o mais próximo possível do valor desejado, e a diferença entre esses valores fornece uma medida de erro de estimação do algoritmo
 - Seja d_j , $j = 1, \dots, n$, a resposta desejada para o objeto j e y_j a resposta estimada (predita) do algoritmo, obtida a partir de uma entrada \mathbf{x}_j apresentada ao algoritmo
 - $e_j = d_j - y_j$ é o sinal de erro observado na saída do sistema para o objeto j

Avaliação de Desempenho

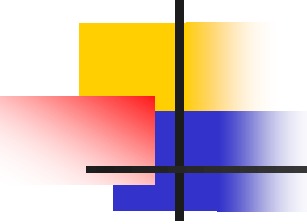
Tabela 6.1 Medidas de desempenho para predição numérica (estimação)

Medida de Desempenho	Fórmula
SSE (SUM SQUARE ERROR) Soma dos erros quadráticos	$SEQ = \sum_{j=1}^n e_j^2$
MSE (MEAN SQUARE ERROR) Erro quadrático médio	$EQM = \frac{1}{n} \sum_{j=1}^n e_j^2$
Raiz do erro quadrático médio	$REQM = \sqrt{\frac{1}{n} \sum_{j=1}^n e_j^2}$
MAE (MEAN ABSOLUTE ERROR) Erro absoluto médio	$EAM = \frac{1}{n} \sum_{j=1}^n e_j $
Erro quadrático relativo	$EQR = \frac{1}{n} \sum_{j=1}^n \frac{e_j^2}{(d_j - \bar{d})^2} \quad , \quad \bar{d} = \frac{1}{n} \sum_{j=1}^n d_j$
Raiz do erro quadrático relativo	$REQR = \sqrt{\frac{1}{n} \sum_{j=1}^n \frac{e_j^2}{(d_j - \bar{d})^2}} \quad , \quad \bar{d} = \frac{1}{n} \sum_{j=1}^n d_j$
Erro absoluto relativo	$EAR = \frac{1}{n} \sum_{j=1}^n \frac{ e_j }{ d_j - \bar{d} }$
Coeficiente de correlação	$\rho = \frac{\sum_{j=1}^n (d_j - \bar{d})(v_j - \bar{y})}{\sqrt{\sum_{j=1}^n (d_j - \bar{d})^2} \cdot \sqrt{\sum_{j=1}^n (v_j - \bar{y})^2}}$



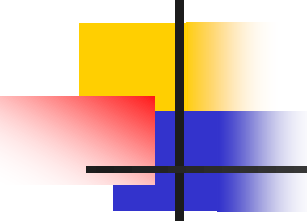
Regressão

- Regressão Linear (Simples, Múltipla)
- Árvore de Decisão (e variações (Random Forest))
- K-NN
- SVM
- ANN
- ...



Árvore de Decisão (e variações (Random Forest))

- A construção de uma árvore de regressão é semelhante à construção de uma árvore de classificação
- A diferença encontra-se na medida a ser utilizada para decidir sobre o melhor particionamento para o conjunto de dados
 - No caso da regressão, busca-se minimizar uma função de custo



Árvore de Decisão (e variações (Random Forest))

- Na referência tem-se a apresentação da seguinte medida

$$SDR(h_A) = sd(\mathbf{D}, \mathbf{y}) - \frac{n_L}{n} \times sd(\mathbf{D}_L, \mathbf{y}) - \frac{n_R}{n} \times sd(\mathbf{D}_R, \mathbf{y})$$

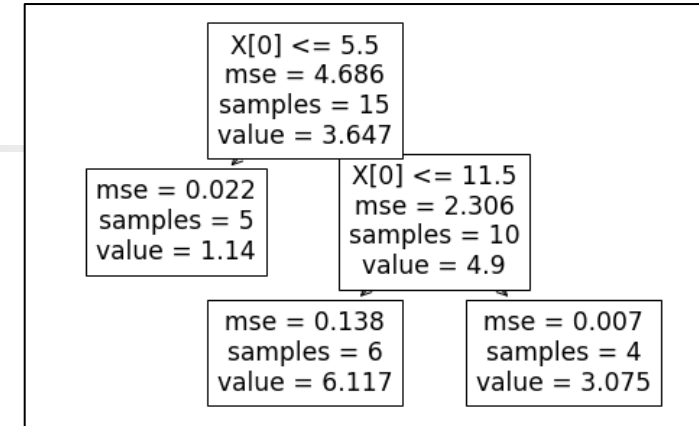
SDR = Standard Deviation Reduction

$$sd(\mathbf{D}, \mathbf{y}) = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}$$

A constante associada às folhas de uma árvore de regressão é a média dos valores do atributo alvo dos exemplos de treinamento que caem na folha

Árvore de Decisão (e variações (Random Forest))

- Outras possibilidades são possíveis



$$SDR(h_A) = sd(\mathbf{D}, \mathbf{y}) - \frac{n_L}{n} \times sd(\mathbf{D}_L, \mathbf{y}) - \frac{n_R}{n} \times sd(\mathbf{D}_R, \mathbf{y})$$

<https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html>

<https://medium.com/analytics-vidhya/regression-trees-decision-tree-for-regression-machine-learning-e4d7525d8047>

$$MSE = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

A constante associada às folhas de uma árvore de regressão é a média dos valores do atributo alvo dos exemplos de treinamento que caem na folha