Design de projetos para criação de modelos preditivos

Fabrício Barth

Julho de 2020

Sumário e Objetivos

- Etapas para construção de modelos preditivos
- Escolha dos dados
- Medidas de erro (para classificação e regressão)
- *Bias*, *Variance*, *overfitting* e *underfitting*

Etapas para construção de modelos preditivos

Etapas para construção de modelos preditivos

- Escolher o conjunto de dados corretos.
- Definir a métrica e a taxa de erro aceitável.
- Dividir os dados em:
  - ⋆ Treinamento.
  - ⋆ Teste.
  - ⋆ Validação (opcional).

- Selecionar atributos que devem formar o conjunto de treinamento.
- Identificar modelos preditivos usando o conjunto de treinamento.

- Aplicar teste sobre o conjunto de treinamento.
- Se não existe conjunto de validação então aplicar o modelo 1x no conjunto de teste.
- Se existe conjunto de validação então aplicar o modelo no conjunto de teste e refinar o modelo.
- Se existe conjunto de validação então aplicar o modelo 1x no conjunto de validação.

Escolha dos dados

Identificando o conjunto de dados corretos

- Em alguns casos é fácil (avaliação de filmes $\rightarrow$ novas avaliações de filmes).
- Em outros pode ser mais difícil (dados genéticos $\rightarrow$ doenças).
- Geralmente, quanto maior a quantidade de dados, melhor são os modelos.
- Conhecer *benchmarks* ajuda!
- **Sempre começamos com dados brutos e precisamos processá-los**.

Medidas de erro

Definição de Erro para problemas de Classificação

Table 1: Conjunto de teste

Exemplo	Classe real	Classe inferida
1	Positivo	Positivo
2	Positivo	Negativo
3	Negativo	Negativo
4	Negativo	Negativo
5	Negativo	Negativo
6	Positivo	Positivo
7	Positivo	Negativo
8	Negativo	Negativo

$$erro(modelo) = \frac{qtd\_incorretos}{qtd\_exemplos} \quad (1)$$

onde:

- $qtd\_exemplos$: quantidade de exemplos do conjunto de teste.
- $qtd\_corretos$: quantidade de exemplos do conjunto de teste incorretamente classificados.

Neste exemplo:

Table 2: Conjunto de teste

Exemplo	Classe real	Classe inferida
1	Positivo	Positivo
2	Positivo	Negativo
3	Negativo	Negativo
4	Negativo	Negativo
5	Negativo	Negativo
6	Positivo	Positivo
7	Positivo	Negativo
8	Negativo	Negativo

$$erro(modelo) = \frac{2}{8} = 0.25 \tag{2}$$

Definição de Verdadeiro e Falso Positivo

- Verdadeiro Positivo = identificado corretamente.
- Falso Positivo = identificado incorretamente.
- Verdadeiro Negativo = rejeitado corretamente.
- Falso Negativo = rejeitado incorretamente.

Exemplo de teste médico:

- Verdadeiro Positivo = Pessoa doente corretamente classificada como doente.
- Falso Positivo = Pessoa saudável incorretamente classificada como doente.
- Verdadeiro Negativo = Pessoa saudável corretamente classificada como saudável.
- Falso Negativo = Pessoa doente incorretamente classificada como saudável.

Matriz de classificação

	Positivo de fato	**Negativo de fato**	**Precisão (Precision)**
Classificados pelo modelo como positivo	Verdadeiro Positivo (VP)	Falso Positivo (FP)	$VP/(VP + FP)$
Classificados pelo modelo como negativo	Falso Negativo (FN)	Verdadeiro Negativo (VN)	$VN/(VN + FN)$
Cobertura (Recall)	$VP/(VP + FN)$	$VN/(FP + VN)$	Acurácia: $(VP + VN)/(FP + FN)$

F1-score

É uma medida harmônica entre *Precision* e *Recall*:

$$F1\_score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3}$$

Curva ROC: exemplo de uso de gráfico na avaliação de modelos

Análise ROC, do inglês *Receiver Operating Characteristic*, é um método gráfico para avaliação, organização e seleção de sistemas de diagnóstico e/ou predição.

Do domínio de Aprendizagem de Máquina, a curva ROC é particularmente útil em casos nos quais existe uma grande desproporção entre as classes ou quando se deve levar em consideração diferentes custos/benefícios para os diferentes erros/acertos de classificação.

Medida de Erro para problemas de Regressão

As medidas de erro mais usadas nesse caso são o raiz quadrada do erro quadrático médio (RMSE - *root mean squared error*) e a distância absoluta média (MAE - *mean absolute error*):

$$RMSE(f) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - f(\vec{x}_i))^2} \tag{4}$$

$$MAE(f) = \frac{1}{n}\sum_{i=1}^{n}|y_i - f(\vec{x}_i)| \tag{5}$$

Quando usar $RMSE$ ou MAE?

- $RMSE$ é uma medida que valoriza os erros extremos, mesmo ocorrendo com pouca frequência. Por exemplo, se para o problema que está sendo analisado errar por 10 é pior que errar duas vezes por 5 então a melhor escolha é $RMSE$. Caso contrário, se errar por 10 é igual que errar duas vezes por 5 então a melhor escolha é MAE.
- Do ponto de vista de interpretação, a melhor escolha é o MAE.

Bias, Variance, overfitting e *underfitting*

Exemplo

- Suponha que você deseja construir um classificador com 5% de erro.
- O erro do seu conjunto de treinamento é de 15%.
- O erro do seu conjunto de treinamento é de 16%.

- O erro associado ao conjunto de treinamento é chamado **bias**.
- O erro associado ao conjunto de teste é chamado **variance**.
- No exemplo anterior temos um **bias** de 15%.
- E uma **variance** de 1% (16 - 15)
- Sendo assim, temos um modelo com alto **bias**.
- Que também é conhecido como **underfitting**.

- Considere um outro exemplo onde:
  - ⋆ erro de treinamento = 1%
  - ⋆ erro de teste = 11%
- Bias = 1%
- Variance = 10% (11 -1)
- Sendo assim, temos um problema de alta **variance**.
- Ou, **overfitting**.

- Considere:
  - ⋆ erro de treinamento $= 0.5\%$
  - ⋆ erro de teste $= 1\%$
- Parabéns! O seu modelo está muito bom!

Material de **consulta**

- Tom Mitchell. Machine Learning, 1997. (Capítulo 5).
- Iah H. Witteh and Eibe Frank. Data Mining, 2000. (Capítulo 5).
- *Prediction study design. Data Analysis Course*. Coursera.org