Validação

Carlos Diego Rodrigues

7 de outubro de 2021

Universidade Federal do Ceará

Validação: avaliando o que foi aprendido

- Questões: treinar, testar, melhorar.
- Predizendo a performance: limites de confiança.
- Separar, validação cruzada, *bootstrap*

Avaliação, a chave do sucesso

- Quão bom é o modelo que foi aprendido?
- Erro nos dados de treinamento não são bons indicativos.
- Solução simples: separar os conjuntos de treinamento e teste.
- Algumas vezes os dados pré-classificados são limitados.

Problemas na avaliação

- Confiança estatística
- Escolha da medida de desempenho
  - Número de classificações corretas
  - Acurácia das estimativas probabilísticas
  - Erro nas predições numéricas
- Custos associados aos erros.

Treinamento e testes

- Medida natural para problemas de classificação: taxa de erro.
  - Sucesso: classe predita corretamente.
  - Erro: classe predita incorretamente.
  - Taxa de erro: proporção de erros sobre todo o conjunto de instâncias.
- Erro de resubstituição: taxa de erro quando o modelo é aplicado ao conjunto de treinamento.
- Erro de resubstituição é otimista!

Treinamento e testes

- Conjunto de teste: conjunto de instâncias independentes que não foram utilizadas na elaboração do classificador.
- Treinamento e teste podem ser de natureza diferente.
  - Exemplo: dados conhecidos sobre os funcionários de uma empresa A podem ser utilizados para fazer inferências sobre uma empresa B.
  - Idealmente porém eles são retirados de dados que possuem uma distribuição comum.
- Em nenhuma das etapas da elaboração do classificador pode haver o uso do conjunto de testes!
- Alguns classificadores vão separar os dados em três: conjunto de treinamento, validação e testes.

Retirando o máximo dos dados

- Uma vez o processo de avaliação esteja completo, todos os dados podem ser usados para construir o classificador final (para usos práticos).
- Em geral, quanto mais dados, melhor o classificador (mas o retorno diminui).
- Quanto maior o conjunto de teste mais precisa é a estimativa do erro.
- Esse procedimento de divisão do conjunto é chamado *holdout*.

Predizendo a performance

- Dada uma taxa de erro obtida de um processo experimental, quão perto ela está da verdadeira taxa de erro?
- Podemos construir uma distribuição amostral da proporção de erros para estimar um intervalo de confiança para o erro.
- Em um processo similar àqueles utilizados na inferência estatística para obter intervalos de confiança para proporções de uma população.

Intervalo de confiança para a taxa de sucesso

- Seja f a taxa de sucesso que obtivemos.
- Seja p a taxa real de sucesso na população.
- Seja c o índice de confiança desejado.

$$P(-z_c \leq X \leq z_c) = c$$

Transformação para a normal padrão

- Transformando para a normal padrão:

$$P\left(-z_c \leq \frac{f-p}{\sqrt{\frac{p(1-p)}{N}}}\right) \leq z_c = c$$

- Resolvendo para p temos para o limite de confiança c:

$$p = \frac{\left(f + \frac{z_c^2}{2N} + z_c\sqrt{\frac{f}{N} - \frac{f^2}{N} + \frac{z^2}{4N^2}}\right)}{1 + \frac{z^2}{N}}$$

Exemplos

- $f = 75\%$, $N = 1000$, $c = 80\%$ (ou seja, $z_c = 1.28$): $[0.691, 0.801]$
- $f = 75\%$, $N = 100$, $c = 80\%$ (ou seja, $z_c = 1.28$): $[0.549, 0.881]$

Como dividir os conjuntos

- Se tivermos apenas um conjunto, como dividir?
- Uma porcentagem para teste e o restante para treinamento.
- Problema: os exemplos podem não ser representativos no treinamento ou teste.
- Solução: estratificação.
- *Holdout* repetido: várias seleções de amostra para teste e treinamento são feitas sobre o mesmo conjunto para efeitos de estabilidade.

Validação cruzada

- Uma validação cruzada de K partes evita a sobreposição.
  - Primeiro passo: dividir o conjunto em K partes iguais.
  - Segundo passo: usar uma das partes para teste e as demais para treinamento.
  - Algoritmo é aplicado em K conjuntos de treinamento diferentes.
- Subconjunto estratificados.
- Erro é a média das K execuções.
- Outros métodos de estimação desse erro.
- Método padrão $K = 10$ estratificado.

Validação cruzada com um-de-fora

- Um-de-fora: técnica particular onde todos os exemplos são usados para treino e apenas um é usado para teste.
- Melhor uso dos dados.
- Não há aleatoriedade.
- Computacionalmente caro pois envolve uma execução para cada instância.
- Não permite estratificação.

Bootstrap

- Validação cruzada não permite amostras com o mesmo elemento mais de uma vez.
- *Bootstrap* permite reposição com elementos do conjunto de treino.
  - Amostragem com reposição em um conjunto com n instâncias para formar um novo conjunto de dados com n instâncias.
  - Use este novo conjunto como treinamento.
  - Use as instâncias do conjunto original que não aparecem nesse conjunto como teste.
- *Bootstrap* 0.632