Machine Learning 02

Generalização

Pau que bate em Chico bate em Francisco

O problema de generalização

- As provas!
- O objetivo não é só se dar bem fora de D
- Dados de treinamento e dados de teste precisam ser independentes, mas da mesma natureza, para assegurar um bom modelo

Erro de generalização

- A grosso modo, é a diferença entre E_{in} e E_{out}
- Enquanto E_{in} pode ser visto com os dados de treinamento, é preciso estimar E_{out}
- Separar pontos de dados não usados em treinamento para teste é uma forma de estimar como a função g se comporta fora de D.
- É uma forma válida se e somente se g e os dados de teste não tiverem nenhuma relação
- Mais dados para teste = menos dados para treinamento :(
- Se os dados de teste são separados aleatoriamente em D, temos várias hipóteses que podem ser g : uma para cada subconjunto de treinamento.

Viés e Variância

- Se o conjunto H é muito simples, podemos nunca encontrar uma hipótese que aproxime bem a função f.
- Se o conjunto H é muito complexo, podemos não conseguir generalizar por causa dos dados de treinamento.
- Erro quadrático e o cálculo de E_{out}
- Viés: erro entre a hipótese média e a função f
- Variância: erro entre a hipótese média e a hipótese escolhida para todo o conjunto D (instabilidade)

Viés e Variância

- A variância diminui quando temos um conjunto D maior.
- Se D for grande o suficiente, só o viés se torna importante
- O viés e a variância são determinados não só por H, mas por A! (Lembra dele?)
- Para melhores modelos de aprendizagem, temos que diminuir um dos componentes de erro sem aumentar significantemente o outro
- E_{out} = viés + variância

Curva de aprendizado

- Curvas que medem o erro (ou a acurácia) de acordo com o tamanho do conjunto de treinamento

Modelo prático

Na prática, a teoria é a mesma

Como implementar um modelo

- Obter os dados!
- Tratamento e escolha dos dados de entrada (features)
- Modelagem
- Testes e avaliação
- Aprimoramento do modelo

Tratar os dados de entrada

- Os dados podem estar incompletos
- Dados categóricos podem ser identificados e tratados
- Dados numéricos podem ser normalizados e ajustados
- Os dados de entrada podem ter relação entre si
- Alguns dados podem ser irrelevantes ou introduzir erros e ruído

Tratar os dados de entrada

- Pré-processamento
- Feature Extraction
- Feature Selection
- Feature Engineering