Machine Learning 03

Overfitting

Elementar, meu caro Watson, roubaram nossa barraca.

Modelo com base nos dados

- Considere a sequência:

  1, 3, 5, 7, ?

- Qual é o valor de ?
- Segundo meu modelo, é 217341.
- Polinômio de 4^o grau com w = [18111 / 2, -90555, 633885 / 2, -452773] e b = 217331.

Complicando coisas simples

- Precisamos falar sobre ***Anatidaefobia***
- Seu modelo não precisa ser tão específico assim
- É um erro de processo: escolher o menor E_{in}

Overfitting

- Acontece quando o seu modelo se apaixona perdidamente pelos dados de treinamento
- E_{in} não é a prioridade para dizer que um modelo é bom.
- Minimizar E_{in} só é útil se E_{out} ~= E_{in}


OVERFITTING

Overfitting

- Pode ser verificado através da curva de aprendizado: E_{in} diminui enquanto E_{out} aumenta
- O máximo de informação sobre o alvo NÃO leva ao melhor resultado, porque você NÃO conhece o alvo...
- A complexidade do modelo deve ser descrita a partir da quantidade e qualidade dos dados

Variações

- Mais dados de entrada, menos overfit
- Mais ruído, mais overfit
- Maior complexidade de f, mais overfit

Não temos controle sobre nada disso.

First things first

- Ruído estocástico
- Ruído determinístico
- $E_{out} = \sigma^2 + \text{viés} + \text{variância}$

A cura

- Diminuir a dimensão de h
- Restringir os pesos de h
- Ter cenários de treinamento diferentes

Regularização

- Métodos mais populares: L1 e L2
- L1: Regularização por módulo
- L2: Regularização por peso quadrado

Regularização demais pode levar a ***underfitting***

Validação cruzada

- Média de vários subconjuntos de treinamento usando vários subconjuntos de teste
- Métodos mais populares: One-vs-All e KFold
- One-vs-All: treinar com N-1 amostras e validar em 1 amostra, N vezes
  - Bom para generalização, indiferente para overfitting
- K-Fold: dividir os dados em K conjuntos de N/K dados, treinar com K-1 conjuntos e validar em 1 conjunto, K vezes
  - Uma boa escolha de K apresenta uma boa generalização e ajuda na redução de overfitting

O modelo Linear

Imagine uma linha imaginária

Modelos lineares de regressão

Conte-me sobre sua infância

Caracterização

- Modelo onde f é uma função com imagem real
- Exemplo: sugerir um limite para o cartão de crédito
- Na prática, temos que lembrar que há ruído: f é uma probabilidade P(y | X)
- A saída é uma combinação linear da entrada

Função de erro

- Least Square error: o treinamento
  - O problema de regressão é minimizar a função de erro do modelo
- Mean Square Error: o teste
  - Diferença entre o valor previsto e o valor original, ao quadrado.

Problemas

- OUT, LIARS!
- O problema pode não ser linear

Se o problema não for linear...

- Linearizamos uai.
- Transformações não-lineares na entrada
- Regressão polinomial

Problemas

- OUT, LIARS!
- ~~O problema pode não ser linear~~
- A maldição da dimensionalidade