Aluno: Ciro B Rosa
No USP: 2320769

MAC5832 - Introdução ao aprendizado de máquina
QT5

P1

Escreva o que você entendeu sobre regularização (no contexto da Lecture 12). Inclua algum comentário sobre a diferença entre usar ou não regularização quando exploramos o espaço de hipóteses.

R1

“Regularização” é uma metodologia, ou ferramenta, que tem como objetivo obter a melhor aproximação g(X) à target function f(X), de modo que esta não produza overfitting.

“Overfitting” ocorre quando uma aproximação qualquer h(X) tenta replicar com exatidão a todas as respotas do training set. Como várias destas respostas podem ter ocorrido devido a ruído (estocástico ou determinístico), a função de aproximação h(X) tenderá a não fazer um bom trabalho de predição de f(X) quando novos vetores X, não existentes no training set, são apresentados. Em outras palavras, quando h(X) procura responder com exatidão a todas as saídas conhecidas de f(X), existe o que se chama de “perda de generalização”, ocasionando imprecisões maiores que desejadas no comportamento de h(X).

Conforme exposto em aula, a regularização é necessária para se diminuir os efeitos de overfitting. Porém, ao ser utilizada, há uma pequena penalidade a ser paga pela falta de exatidão com que h(X) replicará f(X). Esta penalidade pode ser largamente compensada pela escolha adequada do fator λ de regularização, o qual pode reduzir drasticamente os efeitos do ruído na predição.

P2

Na regularização que o prof. Abu-Mostafa chama de weight-decay, o que são C e lambda e qual a relação entre eles? Por que usamos lambda e não C?

R2

O sistema de inequações a ser resolvido para se achar o vetor wreg, a solução regularizada de pesos, é dado por:

$$\text{Minimização de } E_{in}(w) = \frac{1}{N}(Zw - y)^T(Zw - y)$$

$$w^T.w \leq C$$

Onde a matriz Z representa a matriz X remapeada conforme os coeficientes dos polinômios de Lagrange, w é o vetor de pesos, y é o vetor de respostas do sistema, e N é o número de observações do dataset de treinamento.

Desta forma, C é uma constante que impõe um limite máximo aos pesos regularizados (ou à soma dos quadrados de cada um, mais exatamente).
Resolvendo-se o sistema, verifica-se que a minimização de Ein ocorre quando:

$$\nabla E_{in}(w_{reg}) \propto -w_{reg}$$

Ou

$$\nabla E_{in}(w_{reg}) = -k.w_{reg}$$

Onde k é uma constante. Ocorre que o formato de k pode ser escolhido de forma conveniente, sem perda de generalidade por representar uma constante. Especificamente, a escolha a seguir simplifica o formato do resultado da solução.

$$k = 2\frac{\lambda}{N}$$

Sendo que equação a ser minimizada, que dá origem à equação de gradiente

$$\nabla E_{in}(w_{reg}) = -k.w_{reg}$$

é dada por:

$$E_{in}(w) + \frac{\lambda}{N}.w^T w$$

Portanto, C e λ estão relacionados pelo fato de serem duas representações diferentes da mesma restrição aos pesos Wreg.