IBI5031 - Aprendizagem de Máquina para Bioinformática

Segunda Lista de Exercícios

Aluno: Fernanda Midori Abukawa

Novembro 2020

Questão 1

Dimensão VC:

O número da dimensão VC ou d_{vc} dado um conjunto de hipóteses $\mathcal{H}$ é por definição o máximo de pontos que se consegue fragmentar antes do ponto de quebra k, ou seja o máximo número de pontos nos quais aplicam-se todas as dicotomias. Por exemplo no caso do perceptron $2D$, o $k = 4$ e o $d_{vc} = 3$.

Se o d_{vc} for finito, ele garante que a função $g \in \mathcal{H}$ e que g aproxima-se de f. Ou seja, garante a generalização do modelo de aprendizagem, independente do algoritmo escolhido, da distribuição dos dados de entrada e da função alvo.

O d_{vc} também pode ser interpretado como uma medida de graus de liberdade. Por exemplo, a quantidade de parâmetros escolhidos criam esses graus de liberdade. No perceptron $2D$ dado que sua dimensão é 2 ($d = 2$), $k = 4$ e $d_{vc} = 3$, pode-se considerar o valor de d_{vc} como representativo do número de parâmetros ($w0$, $w1$, $w2$) efetivos. Ao acoplar vários perceptrons $2D$, por exemplo 3, o número de parâmetros aumenta para $3 \times 3 = 9$, porém os parâmetros efetivos continuam sendo = 3.

Além disso o número de amostras N se relaciona com d_{vc} da seguinte forma: quanto maior o d_{vc}, maior o N necessário. Dado que ao aumentar os graus de liberdade, precisa-se de mais exemplos para que o E_{in} acompanhe E_{out}. O *rule of thumb* nesse caso é que $N \geq 10d_{vc}$.

Não são todos os modelos de aprendizagem que se consegue relacionar diretamente o número de parâmetros com a d_{vc}, porém consegue-se ter uma ideia de quantos exemplos N são necessários para que ocorra a generalização dado um número de parâmetros.

Balanceamento entre Viés e Variância:

Dado um conjunto de dados $\mathcal{D} \in N$, a função $\bar{g}(x)$ é a média das funções $g(x)$ de cada subconjunto $\mathcal{D}$.

Variância: corresponde à diferença entre a função $g^{\mathcal{D}}(x)$ e $\bar{g}(x)$, interpreta-se como sendo o quão longe está a hipótese $g^{\mathcal{D}}(x)$ da média das hipóteses de todos os subconjuntos $\mathcal{D}$.

Viés: é a diferença entre $\bar{g}(x)$ e $f(x)$, ou seja, o quão longe está a média das hipóteses da função alvo.

O erro fora da amostra esperado é a soma da variância e do viés, por exemplo utilizando o erro quadrático (Eq. 1):

$$\mathbb{E}_{\mathcal{D}}[E_{out}(g^{\mathcal{D}})] = \mathbb{E}_{\mathcal{D}}[(g^{\mathcal{D}}(x) - \bar{g}(x))^2] + (\bar{g}(x) - f(x))^2 \tag{1}$$

Ao aumentar $\mathcal{H}$, aumenta-se a variância e diminui-se o viés. Isso pode ser explicado porque ao aumentar o número de hipóteses, a probabilidade de f se encontrar em $\mathcal{H}$ é maior, porém dado o grande número de g, a diferença de $\bar{g}(x)$ aumenta para cada $g(x)$.

Como se relaciona dimensão VC e Viés-Variância:

No caso da dimensão VC, $E_{out} \leq E_{in} + \Omega$, onde Ω é:

$$\Omega = \sqrt{\frac{8}{N} \quad \ln \frac{4m_H(2N)}{\delta}} \tag{2}$$

e $\delta = 4m_{\mathcal{H}}(2N)e^{\frac{-1}{8}\epsilon^2 N}$, onde $m_{\mathcal{H}}$ é a função de crescimento, ϵ é a diferença entre E_{in} e E_{out} e N é o número de amostras.

Ω é considerado como sendo o erro de generalização (parte vermelha Fig.1), e depende de N,$\mathcal{H}$ e δ. No gráfico da Análise VC da Fig.1, ao aumentar N, diminui-se o erro de generalização, porque quanto mais exemplos mais E_{in} acompanha E_{out}.

Na análise de viés-variância, no gráfico a linha em preto do meio é $\bar{g}(x)$, que é o mesmo valor para qualquer N porque o viés depende do conjunto $\mathcal{H}$. Já a variância (em vermelho) diminui com maior N, porque depende de $\mathcal{D}$.

É importante fazer ambas as análises durante o processo de aprendizagem de máquina. Levando em consideração o tamanho do conjunto $\mathcal{H}$, a complexidade do modelo e o tamanho de N. Lembrando que é necessário conectar a complexidade do modelo com os recursos disponíveis e não à complexidade da função alvo.

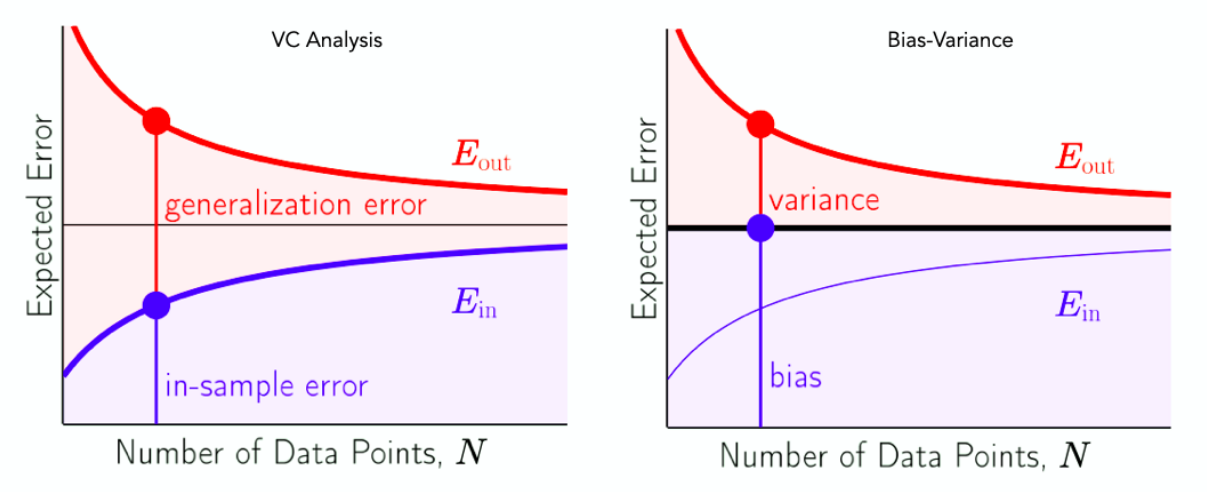

Figure 1: Análise da Dimensão VC e Viés-Variância com relação ao Erro esperado e o número de pontos N, adaptado de Learning from Data, Lecture 8 - Caltech.

Questão 2

O problema do overfitting ocorre quando os dados são ajustados de forma a gerar um valor menor de E_{in} porém que não mais acompanha E_{out}. A função g que vem de dados que sofreram overfitting não se aproxima de f, e portanto não ocorre generalização. Este problema é grave porque gera erros sucessivos durante a análise no aprendizado de máquina.

O overfitting ocorre porque os valores com ruído são ajustados juntamente com os valores sem ruído. Na prática pode ser visualizado computando os valores de E_{in} e E_{out} dados dois conjuntos (treinamento e teste) de uma mesma distribuição e notar que, em um determinado número de épocas enquanto o E_{in} diminui, o E_{out} aumenta (Fig.2), exatamente quando começa a ocorrer o overfitting.

Existem dois tipos de ruídos que compõe o overfitting (Fig.3):

- ruído estocástico: é o erro de cada ponto dado por: $y_n = f(x_n) + \text{ruído}$
- ruído determinístico: é considerado a parte de f que $\mathcal{H}$ não consegue capturar. Esse tipo de ruído depende da complexidade da função alvo e também do conjunto de hipóteses. O ruído determinístico corresponde ao viés da análise viés-variância e é fixo para qualquer $x \in X$.

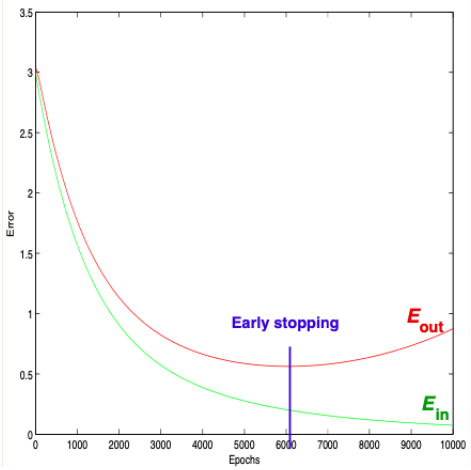

Figure 2: E_{in} e E_{out} dado épocas. O early stopping é utilizado para evitar overfitting. Learning from data, Lecture 11 - Caltech

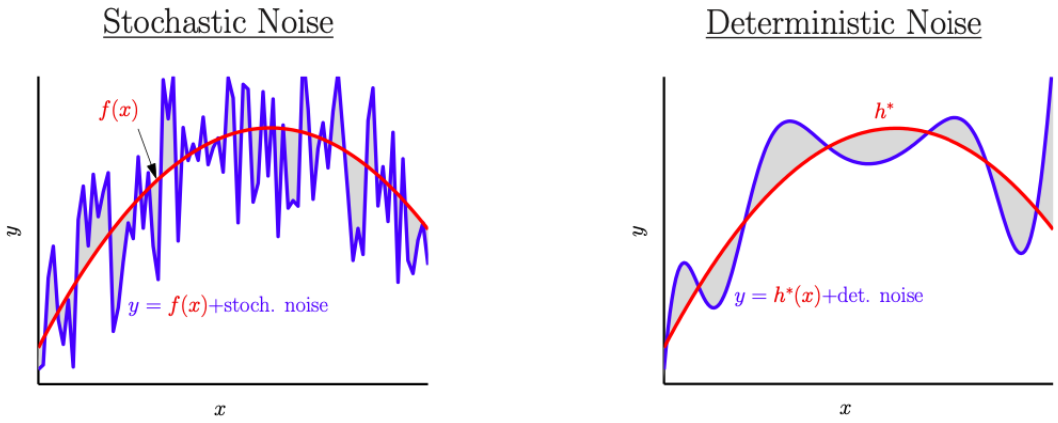

Figure 3: Ruído estocástico e ruído determinístico. Learning from data, Lecture 11 - Caltech

Dado um espaço de hipóteses $\mathcal{H}$ e N número de amostras, o overfitting diminui ao aumentar o N, e aumenta quando ocorre aumento de ruído estocástico e/ou determinístico.

Existem abordagens para mitigar o problema do overfitting, como a regularização e a validação cruzada.

A **regularização** parte do princípio de que o ruído estocástico é de alta frequência (*high frequency*) e o ruído determinístico não é suave (*non-smooth*). Esse método aplica restrições para que os pontos não sejam ajustados perfeitamente a função g, mas que permite "punir" melhor os pontos de ruído. A regularização diminui a variância com um pequeno aumento no viés. A hipótese escolhida será a que for "mais suave", em linhas gerais tendem a escolha de menores pesos. Dessa forma a regularização gera um

erro aumentado E_{aug} (Eq.3), que leva em consideração o regularizador Ω, a taxa de aprendizagem λ e N. E_{aug} é melhor representante/aproximação de E_{out} do que E_{in}.

$$E_{aug} = E_{in}(h) + \frac{\lambda}{N}\Omega(h) \tag{3}$$

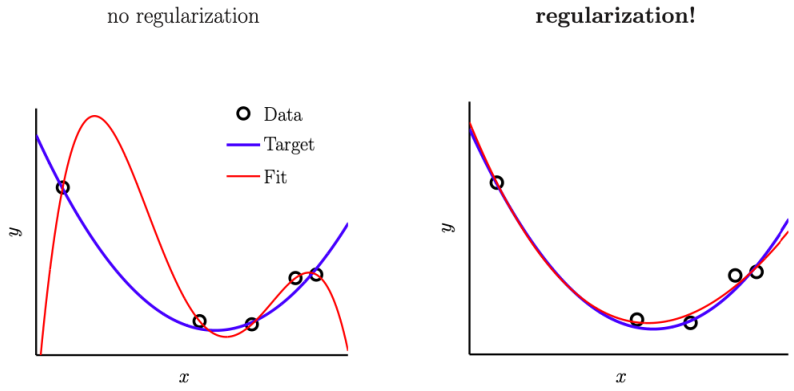

Figure 4: Exemplo de modelos sem regularização (esq.) e com regularização (dir). Learning from data, Lecture 11 - Caltech

Exemplo: ao aplicar LRA, os pesos aplicando-se o regularizador dá-se por:

$$w_{reg} = (X^t X + \lambda I)^{-1} X^t y \tag{4}$$

O método de **validação** é usado como forma de estimar E_{out} e sugere separar o conjunto N em K e $N-K$, utilizando-se $N-K$ para treinar o modelo gerando a função g e K para realizar uma estimativa do erro fora da amostra do modelo treinado (E_{val}).

Com essa técnica, pode-se selecionar o melhor modelo de um conjunto $\mathcal{H}$ verificando para qual modelo E_{val} é menor. Porém deve-se considerar que essa medida de erro é parcialmente enviesada devido ao treinamento prévio que foi realizado no conjunto $N-K$.

Um valor muito grande de K implica um menor número de exemplos para treinamento, o que aumenta E_{in}, porém E_{val} se aproxima melhor de E_{out}. O *rule of thumb* do valor de K é $\frac{N}{5}$.

Exemplo: aplicar *k-fold* para um $N > 100$, com $\frac{N}{K}$ iterações e $N-K$ pontos para treinamento. O *rule of thumb* nesse caso é $K = \frac{N}{10}$.

Definições Importantes para Questão 1

Não considerar o número de palavras dessa seção. Como a questão 1 tem muitos termos importantes, aqui vai uma pequena lista de definições.

Dicotomias: dado que a princípio o espaço de hipóteses é infinito, pode-se utilizar a estratégia de aplicar as chamadas dicotomias para reduzir o espaço de hipóteses $\mathcal{H}$. Isso leva em consideração que os eventos ruins são muito sobrepostos e que portanto, muitas das hipóteses de $\mathcal{H}$ podem ser reduzidas a apenas uma hipótese. Dessa forma, em vez de considerar todas as hipóteses do conjunto, restringe-se para apenas a amostra N e determina-se dicotomias, as n possibilidades de classificar os diferentes pontos de diferentes formas. Essas n possibilidades é um número finito.

Função de crescimento: dado N pontos, é o valor máximo de dicotomias possíveis. Por definição o máximo dessa função é sempre 2^N.

Ponto de Quebra: se existir escolha de k pontos nos quais consegue-se gerar todas as possíveis dicotomias, então k é o break point de $\mathcal{H}$. Por exemplo, em um perceptron $2D$, o $k = 4$.