Universidade Federal de Viçosa
Centro de Ciências Exatas e Tecnológicas
Departamento de Informática

Avaliação de Models de Aprendizagem
Levi H. S. Lelis
levi.lelis@ufv.br

Polarização e Variância de um Modelo

Considere um problema de regressão onde o objetivo é prever os valores de uma função $f(x)$.

Considere um algoritmo que induz hipóteses g_D a partir de amostras D de tamanho m. O erro quadrático esperado dessas hipóteses para conjuntos D para instâncias $x \notin D$ é dado por,

$$E_D[(f(x) - g_D(x))^2] \,.$$

Com um pouco de álgebra é possível decompor a equação acima em duas partes.

$$E_D[(f(x) - g_D(x))^2] = (f(x) - E_D[g_D(x)])^2 + E_D[(g_D(x) - E_D[g_D(x)])^2]$$

O primeiro termo é conhecido como o termo de **polarização** e o segundo como o termo de **variância**.

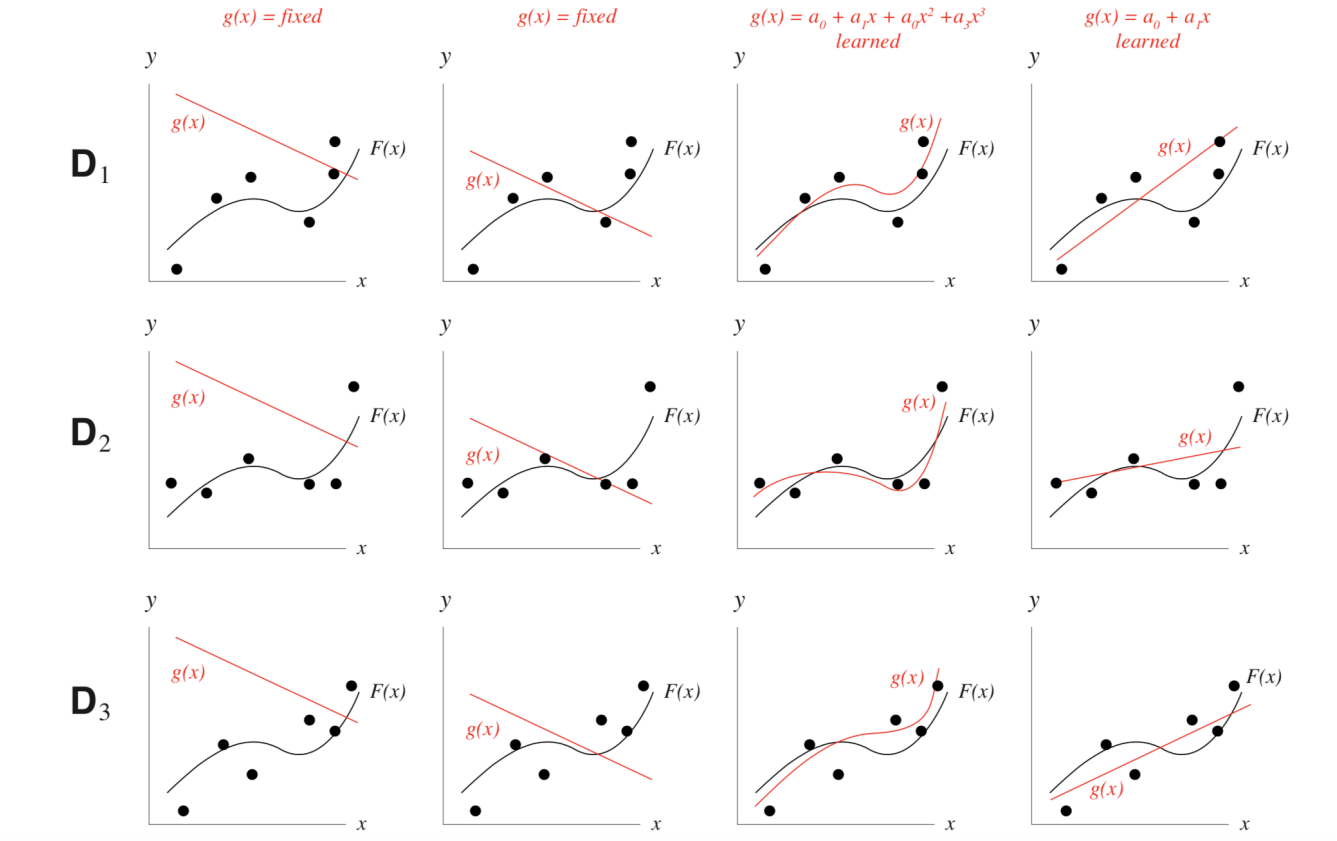
Polarização: Mede o erro esperado da hipótese para diferentes conjuntos D na instância x.

Variância: Mede o erro causado pelas variações de g_D para diferentes conjuntos D.

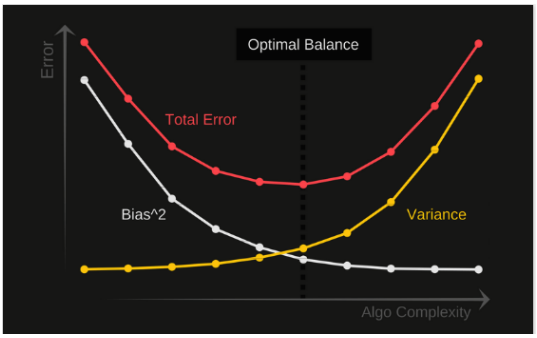
A figura abaixo exemplifica os conceitos de polarização e variância.[1]

- Os pontos de cada um dos gráficos são extraídos (com ruído) de uma equação cúbica $F(x)$.
- D_1, D_2 e D_3 representam diferentes conjuntos de treinamento.
- Nas primeiras duas colunas as hipóteses g são fixas. Por isso o termo da variância das duas são iguais a zero: mesmo trocando o conjunto de treinamento D as hipóteses permanecem inalteradas.
- A polarização de g da segunda coluna é menor que a polarização de g da primeira coluna, pois a reta melhor aproxima a equação $F(x)$.
- A polarização de g da terceira coluna é menor que das duas primeiras, já que g se parece mais com F. A variância de g da terceira é maior que todas outras, já que pequenas modificações em D fazem g variar.
- A polarização de g da última coluna é maior que da coluna anterior, já que a equação linear g é mais distante da equação sendo aproximada. A variância, por outro lado, é menor na última coluna que na coluna anterior, visto que a reta varia menos com as mudanças em D.
- Quando $m \to \infty$, apenas o termo de polarização da terceira coluna vai para um número pequeno (não vai a zero devido ao ruído dos dados); o termo de variância de todos os modelos vai a zero quando $m \to \infty$.

[1] Extraído de *Pattern Classification* de Richard Duda, Peter Hart e David Stork.



De uma forma geral, para conjuntos de treinamento de tamanho m, modelos muito simples (e.g., Regressão Logística ou RNA com poucos neurônios) tendem a ter um erro de polarização grande, já que não conseguem modelar adequadamente a função f. No entanto, ao aumentarmos a complexidade do modelo, o termo de polarização tende a diminuir e a de variância aumentar, como mostrado no esquema abaixo.



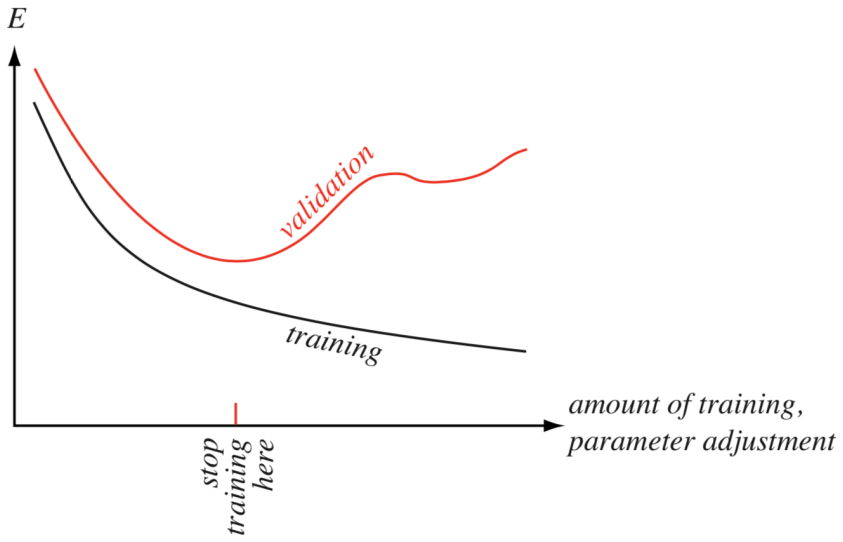
Reparem que o esquema acima funciona para valores fixos de m. Quando aumentamos o tamanho do conjunto de treinamento e o modelo possui complexidade suficiente para aprender f, conseguimos reduzir tanto a polarização quanto a variância.

Os termos de polarização e variância dependem de valores desconhecidos na prática. Como verificar se os

nossos modelos estão encontrando o ponto certo entre polarização e variância?

Conjuntos de Treinamento, Validação e Teste

Separamos D em subconjunto disjuntos de treinamento D_t e D_v. Assim, o modelo pode ser treinado em D_t e o valor do **erro de generalização** avaliado em D_v. Se durante o treinamento o erro em D_t estiver diminuindo mas o erro em D_v estiver aumentando, nós provavelmente cruzamos o ponto ótimo entre polarização e variância. O treinamento deve ser interrompido assim que o erro de validação começar a aumentar, como ilustrado na figura abaixo.[2]



É comum também dividir o conjunto D em três subconjuntos: treinamento (D_t), validação (D_v) e teste (D_e). Esse tipo de divisão é útil se quisermos usar o conjunto de validação para estimar o erro de generalização do modelo para diferentes hiper-parâmetros do modelo.

Para escolher hiperparâmetros do modelo, faça:

1. Para cada hiperparâmetro p:
   (a) Treine modelo com p em D_t
   (b) Avalie modelo treinado em D_v
2. Reporte os resultados em D_e do modelo usando o valor p com melhores resultados em D_v.

Exemplo: Podemos treinar várias versões de uma RNA com diferentes arquiteturas e observar qual tem o menor erro no conjunto de validação. Uma vez escolhida a arquitetura, o erro reportado do sistema é aquele obtido no conjunto de teste.

[2]Extraído de *Pattern Classification* de Richard Duda, Peter Hart e David Stork.

Validação Cruzada

n-partições Uma generalização do processo de uso de um conjunto de validação é o de validação cruzada. A ideia é separar o conjunto D (ou o conjunto $D \setminus D_e$, se quisermos trabalhar com conjunto de teste) em n partes. Então, uma aproximação ao **erro de generalização** é dado pela média dos erros para os n possíveis conjuntos de validação. Isto é, o modelo é treinado com os dados de $n-1$ partes e avaliação na parte deixada de fora. Esse processo é repetido para cada possível combinação de partes.

Deixe-uma-de-fora Em um caso extremo, o valor de n pode ser igual a m e o conjunto de validação é composto de apenas 1 instância. Esse esquema permite conjuntos de treinamentos maiores, mas aumenta o custo computacional do teste.

Matriz de Confusão

Suponha que um sistema tenha 99% de precisão no conjunto de teste em um problema de classificação binária. Esse é um bom sistema? Depende, pois as classes podem ser bastante desbalanceadas como 99.99% da classe negativa e 0.01% da classe positiva. Uma forma de reportar resultados é através de uma matriz de confusão.

	Predita Positiva	**Predita Negativa**
Real Positiva	Verdadeiro Positivo (VP)	Falso Negativo (FN)
Real Negativa	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Para o problema de diagnóstico de meningite, em que a maioria dos casos são negativos, poderíamos ter uma matriz de confusão como a exibida abaixo. Em termos de porcentagem de objetos classificados corretamente,

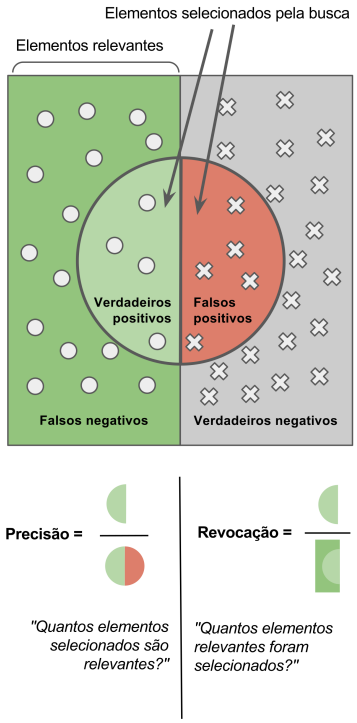
	Predita Positiva	Predita Negativa
Real Positiva	10	2
Real Negativa	100	50000

o problema assim é bastante preciso: $\frac{50010}{50112} = 99.7\%$. No entanto, 2 de 10 pacientes da classe positiva (16.66%) não receberiam o tratamento adequado, o que pode ser considerado um número muito alto para que o sistema seja utilizado em um cenário real.

Precisão e Revocação (Precision and Recall)

Uma outra forma de apresentar resultados similares ao da matriz de confusão são os conceitos de precisão e revocação, da literatura de recuperação de informação.

Dada uma consulta em uma máquina de busca, podemos avaliar o sistema através da fração entre o número de documentos relevantes retornados pelo número de documentos existentes que são relevantes para a busca (revocação). E através da fração entre o número de documentos retornados que são relevantes pelo total de documentos retornados.



A figura ao lado,[3] ilustra os conceitos de precisão e revocação. No caso da meningite, a precisão é dada por,

$$\frac{10}{110} = 0.09\,.$$

e a revocação por,

$$\frac{10}{12} = 0.83\,.$$

Em um problema médica como esse, talvez uma precisão baixa não seja tão problemática quanto uma revocação ainda longe de perfeita. No entanto, os números de precisão e revocação fornecem mais informação que a simples porcentagem de objetos classificados corretamente.

[3] Extraída do Wikipédia.