Complejidad de modelos: Sesgo y Varianza

17 de abril de 2008

Notas de clase.
Rolando Beltrán A

Las medidas de sesgo y varianza son útiles para los modeladores en tanto que ayudan a regular la complejidad del modelo final. Estas medidas se relacionan con la capacidad de ajuste y generalización de un modelo. Cuando se logra un gran ajuste, la diferencia entre los datos reales y la estimación del modelo es pequeña, en este caso el sesgo también es pequeño, pero estos buenos resultados de ajuste, van de la mano con el aumento en la complejidad del modelo, cuando se aumenta la complejidad del modelo este se vuelve sensible a pequeñas variaciones en los datos de entrada, fluctuando en función de estos, es así cuando la varianza aumenta. Esta claro que en Aprendizaje de Maquina se busca crear modelos que ofrezcan dos características esenciales: ajuste a los datos y generalización. Esto acentúa la necesidad de encontrar un balance entre sesgo y varianza, o visto de otra forma, entre error y complejidad, a continuación se describen algunos conceptos importantes referentes al tema.

La utilidad del análisis del sesgo y la varianza pueden ser introducidos analizando el problema de la regresión:

Sea:

$\chi_i = \{x^t, r^t\} \quad i = 1....M$ una muestra i de observaciones, conformada por datos de entrada y su respectiva salida.

$r = f(.)$ una función de salida para el universo de los datos de entrada

$r(x) = f(x) + \epsilon$ la función para un conjunto de datos χ

con

$f(x)$una función continua

$\epsilon \sim N(0, \sigma^2)$

esto es, $r(x)$ esta definida por una función $f(x)$ y ruido ϵ, lo que describe el hecho que los datos reales no son perfectos. Esta claro que en problemas del mundo real raramente se conoce la función $f(x)$ correspondiente a los datos observados.

La tarea es encontrar un estimador $g(.)$que explique los datos observados.

Es posible evaluar la efectividad del estimador para una muestra especifica, utilizando la medida de error cuadrático esperado sobre un conjunto dado:

$$E[(r - g(x))^2|x] = \underbrace{E[(r - E[r(x)])^2|x]}_{ruido} + \underbrace{(E[r(x)] - g(x))^2}_{error\,cuadratico}$$

el primer termino llamado ruido, es independiente del estimador g(.), solo depende de ϵ.

dado que:

$E[r(x)] = E[f(x) + \epsilon]$
$= E[f(x)] + E[\epsilon]$
se tiene que:
$r - E[r(x)] = \epsilon$

esto representa el error inherente a la estimación, el cual es irreducible independientemente del estimador que se utilice.

El segundo termino mide que tanto se desvía el estimador $g(x)$ del valor esperado de la función que genera los datos de salida. Se observa que este termino depende de el estimador $g(x)$ y del conjunto de entrenamiento. Por supuesto el estimador servirá bien para algunos conjuntos de datos y de manera menos efectiva para otros. En el caso de la regresión habrá un $g_i(x)$óptimo para cada conjunto de datos, ver Figura 1.

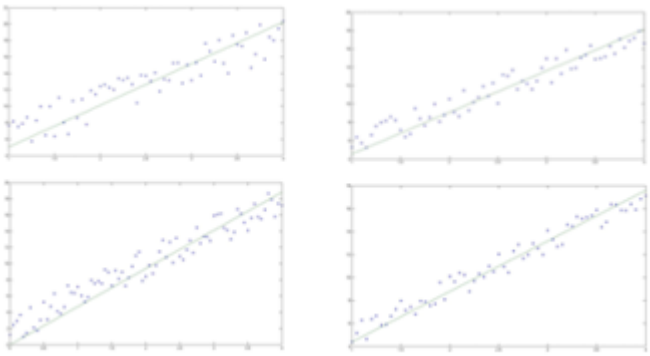

Figura1. Cuatro conjuntos de datos con $r = 4x + ruido$. La regresión crea un estimador nuevo para cada conjunto de datos.

Para evaluar un estimador en necesario probarlo sobre un numero M de conjuntos de datos (cada uno con tamaño N), haciendo esto se tiene:

$$E_\chi[(E[r(x)] - g(x))^2|x] = \underbrace{(E[r(x)] - E_\chi[g(x)])^2}_{----sesgo^2} + \underbrace{E_\chi[(g(x) - E_\chi[g(x)])^2]}_{varianza}$$

El primer termino es llamado sesgo, mide la diferencia entre el valor real y el valor esperado, mientras el segundo llamado varianza mide la fluctuación de $g(x)$ alrededor de el valor esperado $E[g(x)]$, sobre los ejemplos.

El valor medio, $E[g(x)]$, puede ser estimado como el promedio de $g_i(\cdot)$, esto es:

$\bar{g}(x) = \frac{1}{M}\sum_{i=1}^{M} g_i(x)$

Para el caso del sesgo y la varianza tenemos:

Sesgo$^2(g) = \frac{1}{N}\sum_t [\bar{g}(x^t) - f(x^t)]^2$

Varianza$(g) = \frac{1}{NM}\sum_t \sum_i [g_i(x^t) - \bar{g}(x^t)]^2$

Como ejemplo especifico supongamos que conocemos la función $f(x) = 4sen(\frac{3}{2}x)$ que define el valor de salida para los conjuntos de datos.

Seleccionemos dos estimadores con diferente complejidad, ver Figura 2.

El primero :

$g(x) = 5$

es evidente que la varianza es nula pues se trata de una constante, pero también es obvio que el sesgo es bastante alto por que la estimación no toma en cuenta los datos, la única posibilidad de que sea un buen estimador es que $f(x)$ se una función parecida a 5.

La segunda estimación es:

$g(x) = \sum_t r^t/N$

es decir un promedio de los valores de salida de la muestra, esta estimación es mucho mejor a la anterior porque toma en cuenta las observaciones, con lo que se reduce el sesgo, pero aumenta la varianza por su valor cambia de acuerdo al conjunto de datos elegido.

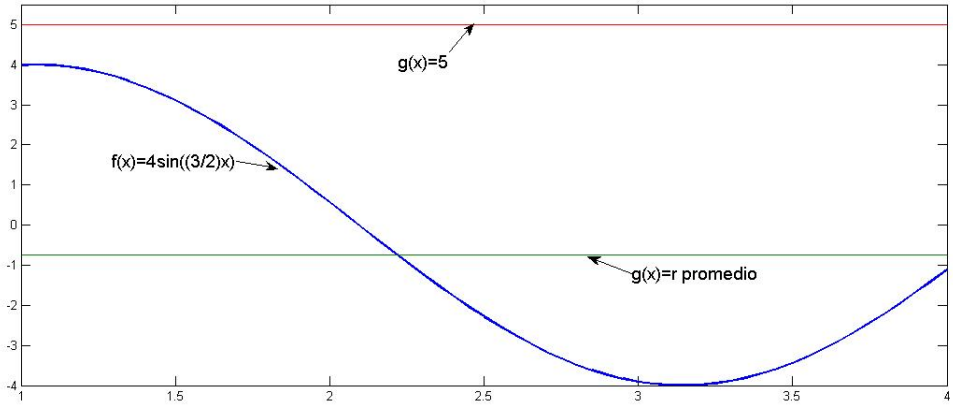

Figura 2. $f(x) = 4sen(\frac{3}{2}x)$ es la función real. El estimador $g(x)$tiene un gran sesgo pues es independiente de los datos, pero su varianza es nula. El estimador $g(x) = \sum_t r^t/N$ tiene menor sesgo pero fluctúa de acuerdo al conjunto de datos.

Otro ejemplo es la utilización de estimadores polinomiales, la figura 3 ilustra este caso, los modelos complejos, en este ejemplo los polinomios de mayor orden ajustan mucho mejor la función real, disminuyendo el sesgo, sin embargo fluctúan mucho de acuerdo a los conjuntos de datos que son utilizados

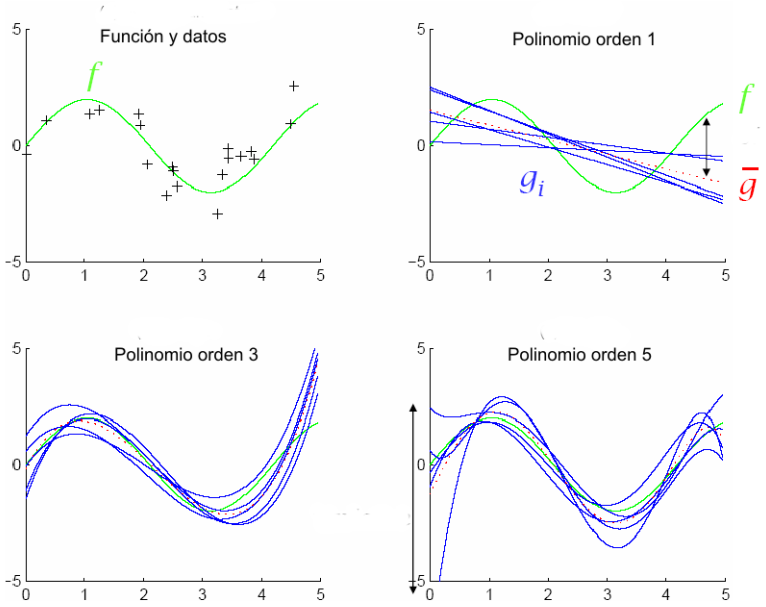

Figura 3. La curva verde representa la función real, los polinomios de estimación están en azul.Los estimadores polinomiales con grado mayor ajustan mejor y presentan menor sesgo, pero fluctúan mucho de acuerdo a los datos de entrada. Es decir a mayor complejidad mayor varianza.

Los ejemplos anteriores describen el problema llamado, dilema del sesgo/varianza, el cual es inherente al proceso de modelado en general, cuando existe gran sesgo, el modelo sugerido esta lejano a la solución, es decir hay sub-ajuste (underfitting), cuando se presenta una varianza considerable, el modelo se ajusto incluso al ruido, lo que se llama sobre-ajuste (overfitting). Es por tanto necesario encontrar una balance entre sesgo y varianza.

Procedimientos de ajuste de complejidad. Existen varios criterios y procedimientos para lograr el balance entre complejidad y ajuste de un modelo.

Para describir el ***cross validation***, suponga que tiene un conjunto de datos con 1000 ejemplos, se deben formar dos subconjuntos, uno para entrenamiento y otro para pruebas, digamos 700-300. Sobre el subconjunto de entrenamiento se realiza una subdivisión otra vez en dos grupos, puede ser 630-70, uno para entrenar los modelos candidatos con diferente complejidad, y otro para validar y hacer ajuste en los hiper-parámetro. Se grafican juntos el error de entrenamiento y el de validación, contra la complejidad del modelo (Figura 4), el error comienza disminuyendo con el aumento de la complejidad, pero después de aumentarla mucho, se observa que el error no mejora demasiado, e incluso empeora, el punto óptimo de complejidad corresponde al codo de la gráfica del error de

validación. El conjunto llamado de prueba no debe ser utilizado para ajustar hiper parámetros, solo se debe usar para reportar resultados, por eso también es llamado publication set.

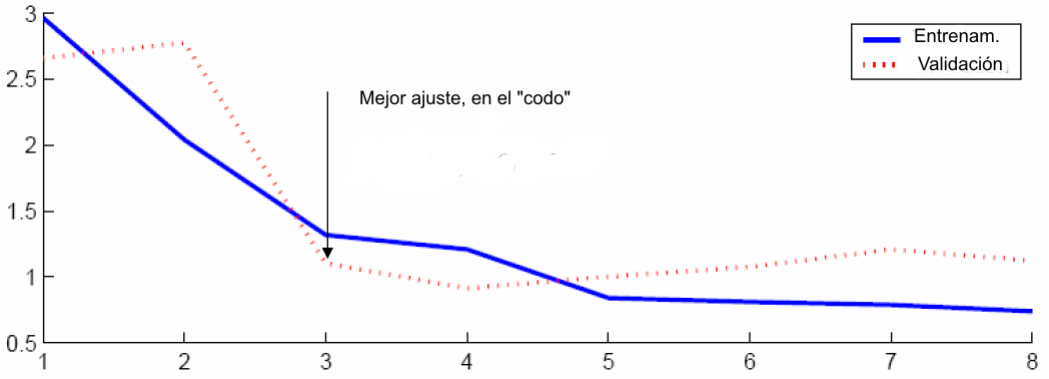

Figura 4. Gráfica de error cuadrático de entrenamiento y de validación contra complejidad del modelo. El codo en el error de validación representa el punto óptimo de complejidad

Entre otros modelos de selección tenemos:

Regularización, el cual describe una función de error extendida

E' =error en los datos + λ·complejidad del modelo

minimizando esta función se busca aumentar el ajuste a los datos y a la vez castigar los modelos complejos, cuando λes muy alto, solo es posible la elección de modelos simples.

Minimum Description length, se basa en la complejidad de Kolmogorov, y busca encontrar la descripción más corta para un conjunto de datos. Entre todos los modelos que describen los datos se selcciona el que haga la descripción más corta. Por ejemplo si se tiene una secuencia de ceros, una forma sencilla de describirla es colocando un cero y el tamaño de la secuencia.

Selección Bayesiana, se utiliza un modelo bayesiano en el que se introduce una probabilidad a priori asignada por medio del conocimiento subjetivo sobre los modelos candidatos, otorgando mayor probabilidad a los modelos más simples. La probabilidad a posteriori determina que modelo es el elegido, y esta definida por el conocimiento subjetivo que describe la probabilidad a priori y el conocimiento objetivo sacado de los datos.