Teoria de la decision

Matias Bajac

2024-09-21

Recapitulando…

Donde estamoss?!!!

$$Y \ = \ f(X) \ + \ \epsilon$$

- **Objetivo**: estimar f()

-tipos de problemas: regresion vs clasificacion

- Porque estimar f, prediccion vs inferencia
- Compromiso flexibilidad en ajuste vs interpretabilidad de resultados

Problema de regresión

Objetivo: construir un modelo que permita **predecir** las ventas en nuevos mercados, y entender en que medios es conveniente invertir

2 aproximaciones

- Vecino mas cercano
- Modelo lineal

##Modelo vecino mas cercano##

Comenzamos usando solamente TV como variable explicativa o independiente

$$Sales_i \ = \ f(TV_i) \ + \ \epsilon) =$$
$$E(Y_i|X_i = x_0) + \epsilon$$

Estimamos f(x) con el **promedio local** de los K puntos mas cercanos a x (k es el numero de vecinos)

$$\hat{f}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i$$

$$N_x = \{x_i \in D : d_i \leq d_k\}$$

Vecino mas cercano

Donde:

- $d_i^2 = d_2^2 = \sum_j (x = x_{ij})^2$ distancia Euclidiana

Cantidad de vecinos

El numero de vecinos k, tiene un efecto importante en el resultado.

- Pocos vecinos: resultados muy pegados a los datos, **overfitting**
- Muchos vecinos: resultado poco senisble, promedio global.

Modelo con todas las explicativas

- El metodo es el mismo usando las 3 variables

-Antes de comenzar hay que elegir el valor de k, cantidad de vecinos

- Para evaular el modelo hay que calcular algun tipo de error

Evaluacion de modelos

- No hay un metodo estadistico mejores a otros para todos los posibles problemas.
- Algunos metodos funcionan bien en algunos problemas y no en otros.
- Hay que decidir para cada problema cual metodo prodduce mejor resultado
- Seleccionar la mejor solucion es uno de los puntos mas desafiantes del aprendizaje estadistico

Trade off - sesgo varianza

Como vimos antes. f se estima a traves de los datos, ya que no estamos suponiendo un modelo para el mismo.

Para evaular la performance de un metodo en un conjunto de datos se necesita contar con formas de medir que tan cercanas se encuentran las predicciones de los datos observados.

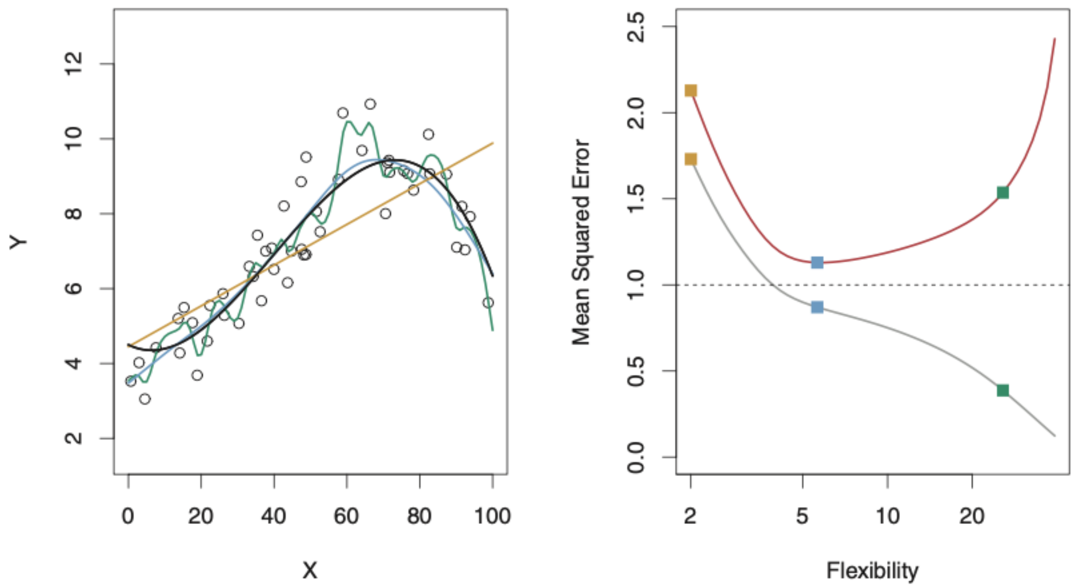
- Conjunto de entrenamiento (train)

$L = \{(x_i, y_i)\}^n \ = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$

-Conjunto de prueba(test): $T = \{(x_0, y_0)\}$

Performance del modelo

- **Regresion** : ECM
- con los datos de entrenamiento: $MSE_{train} = \frac{1}{n}\sum_{i=1}^{i=n}(y_i - \hat{f}(x))^2$
- con los datos de testeo: $MSE_{test} = AVE(y_0 - \hat{f}(x_0))^2$
- **Clasificacion** ERROR DE CLASIFICACION
- con los datos de entrenamiento: $error_{train} = \frac{1}{n}\sum_{i=1}^{i=n} I(y \neq \hat{f}(x_0))$



Datos simulados a partir de una cierta funcion f(grafico en negro). Se ajusta recta de regresion y dos modelos flexibles (splines)

- $\uparrow\ flexibilidad\ \rightarrow \downarrow\ MSE_{train}\ pero\ no\ el\ MSE_{test}\ (forma\ de\ U)$

Esto de que ante mayor felxibilidad menor error cuadratico medio de entrenamiento es porque el modelo ajusta mejor a los datos que ya conoce.

El MSE de test empieza a aumentar en modelos flexibles debido a que los datos de testeo no generalizan bien los nuevos datos.

- $Si\ tenemos\ MSE_{train}\ bajo\ pero\ MSE_{test}\ alto\ hay\ overfitting$

hay que tener cuidado porque a veces un modelo muy flexible puede generar overfitting

Compromiso sesgo - varianza

La froma de U en las curvas MSE_{test} es el resultado del compromiso sesgo varianza.

$$E(y_0 - \hat{f}(x_0))^2 = V(\hat{f}(x_0)) + bias(\hat{f}(x_0))^2 + V(\epsilon)$$

Para minimizar el error test esperado necesitamos f que tenga simultaneamente varianza y sesgo pequeños.

- La vatianza de un modelo de AA mide cuanto varia $\hat{f}$ al modificar el conjunto de entrenamiento, es decir es una medida de la estabilidad tecnica (en general los metodos mas flexibles tienen mayor varianza)
- El sesgo de un modelo AA cuantifica el error de eleccion del modelo (en general los metodos mas flexibles tienen menos sesgo)

El objetivo consiste en encontrar una funcion que minimice el riesgo de predecir mal, para ello se define una funcion de perdida L(X,f(X),Y) y se busca f* entre todas las funciones de una cierta clase C, que haga minimo el valor esperado de L

$$E(L(X, f(X), Y))$$

que llamamos riesgo, en la practica buscamos aquella funcion $\hat{f}$ que minimiza el riesgo empirico

$$\frac{1}{n}\sum_{i=1}^{i=n} L(X_i, f(X_i), Y_i)$$

Errores

Si f* es el mejor entre todos los predictores posibles, f** el mejor de los predictores posibles en una cierta clase de funiones C y $\hat{f}$ el predictor que usamos en la practica tenemos dos tipos de error:

- Error de modelizacion f - f**: depende de la eleccion de la clase C, si consideramos como la familia de todas las funciones posibles, tendremos overfitting.
- Error de estimacion f** -$\hat{f}$: es un error estadistico, si el tamanio n de la muestra es grande, bajo cierta hipotesis sobre la clase C, se cumple que $\hat{f}$ converge a f**