Aprendizaje Supervisado - Tema 1

Aprendizaje Estadístico

Lorena Romero Mateo - 77857300T

Email: lorena.romerom@um.es

Índice

1. Problema de aprendizaje supervisado

Punto inicial: Medición del resultado Y (también llamada variable **dependiente**, respuesta, objetivo).

- Vector de p mediciones predictoras X (también llamadas entradas, regresores, covariables, características, variables independientes).
- En el problema de regresión, Y es cuantitativa (por ejemplo, precio, presión arterial).
- En el problema de clasificación, Y toma valores en un conjunto finito y no ordenado (sobrevivió/murió, dígito 0-9, clase de cáncer de muestra de tejido).
- Tenemos datos de entrenamiento $(x_1, y_1), \ldots, (x_N, y_N)$. Estas son observaciones (ejemplos, instancias) de estas mediciones.

Training data no es lo mismo que muestra, tenemos una muestra de la población y tenemos que sustraer varios datos para el conjunto de entrenamiento y el de testeo.

1.1. Modelos paramétricos y estructurados

El modelo linear es un ejemplo importante de un modelo paramétrico:

$$f_L(X) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p \tag{1}$$

- Un modelo lineal se especifica en términos de $p + 1$ parámetros $\beta_0, \beta_1, \ldots, \beta_p$.
- Estimamos los parámetros ajustando el modelo a los datos de entrenamiento.
- Aunque casi nunca es correcto, un modelo lineal a menudo sirve como una buena y comprensible aproximación a la función verdadera desconocida $f(X)$.

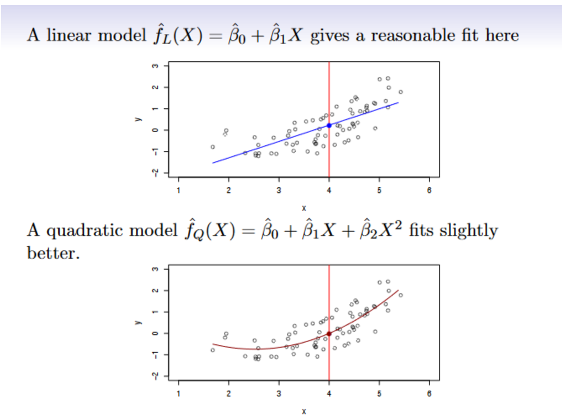


Figura 1: Ajuste regresión lineal vs cuadrática

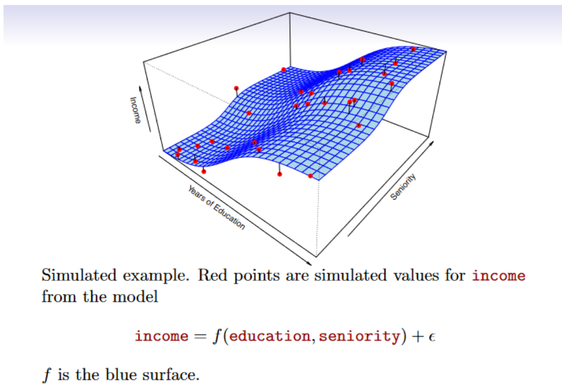


Figura 2: Ejemplo de ajuste

Imagen que aparece a la izquierda:

Los gorritos que aparecen sobre los parámetros indican que son estimaciones de los mismos.

Ventajas del modelo lineal: se interpreta muy fácil y enseguida nos damos cuenta de si va a cuadrar o no.

Se llama recta de regresión debido a que esperamos una recta que modeliza como la media de Y más algo, tiende a acercarse a la media de la variable.

Imagen que aparece a la derecha:

El epsilon es para representar el ruido en los datos.

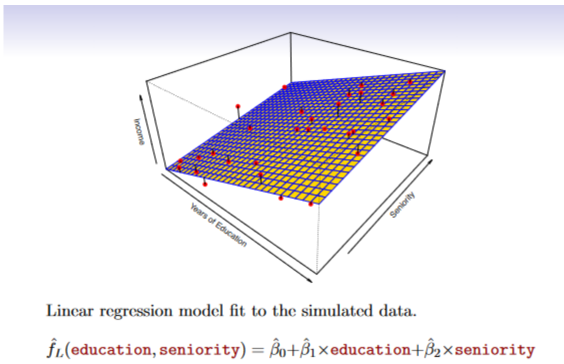


Figura 3: Ajuste regresión lineal

El siguiente tema no lo vamos a dar seguramente:

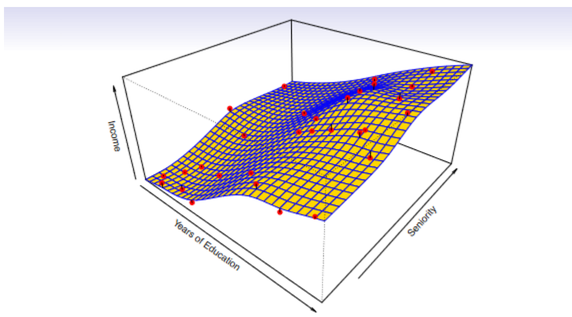


Figura 4: Thin-plate-spline

Modelo de regresión más flexible $\hat{f}_S$ (educación, antigüedad) ajustado a los datos simulados. Aquí utilizamos una técnica llamada **thin-plate spline** para ajustar una superficie flexible. Controlamos la rugosidad del ajuste (capítulo 7). Se calcula por interpolación, el modelo final tiene mucho sesgo. No es un modelo paramétrico.

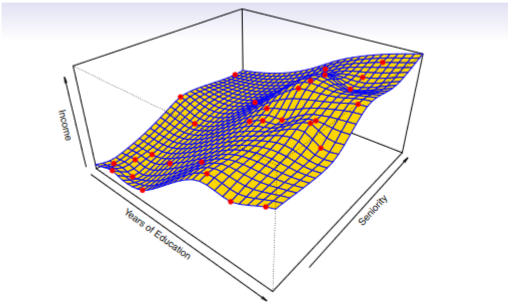


Figura 5: Overfitting

Modelo de regresión spline aún más flexible $\hat{f}_S$ (educación, antigüedad) ajustado a los datos simulados. ¡Aquí el modelo ajustado no comete errores en los datos de entrenamiento! También conocido como overfitting.

1.2. Algunas compensaciones

Precisión de la predicción versus interpretabilidad.

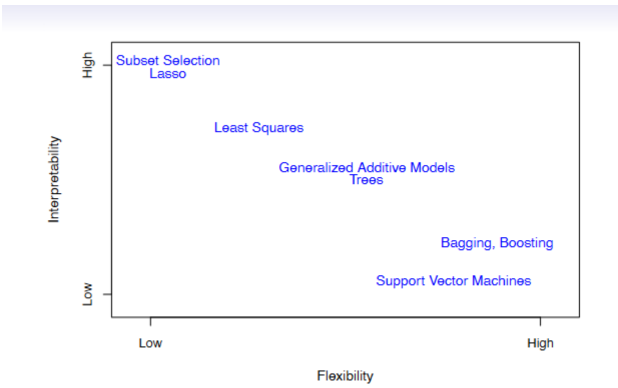
- Los modelos lineales son fáciles de interpretar; los splines thin-plate no lo son.

Buen ajuste versus sobreajuste o subajuste.

- ¿Cómo sabemos cuándo el ajuste es el adecuado?

Parsimonia versus caja negra.

- A menudo preferimos un modelo más simple que involucre menos variables sobre un predictor de caja negra que las involucre todas.
- Si el experto no entiende el modelo no va a ser útil.



2. Evaluación de la Precisión del Modelo

2.1. Bondad de ajuste

Supongamos que ajustamos un modelo $\hat{f}(x)$ a algunos datos de entrenamiento $Tr = \{(x_i, y_i)\}_{i=1}^{N}$, y deseamos ver qué tan bien se desempeña.

- Podríamos calcular el error cuadrático medio de predicción sobre Tr:

$$\text{MSE}_{Tr} = \text{Ave}_{i \in Tr}[(y_i - \hat{f}(x_i))^2]$$

  Esto puede estar sesgado hacia modelos con sobreajuste.

- En su lugar, deberíamos, si es posible, calcularlo usando nuevos datos de prueba $Te = \{(x_i, y_i)\}_{i=1}^{M}$:

$$\text{MSE}_{Te} = \text{Ave}_{i \in Te}[(y_i - \hat{f}(x_i))^2]$$

Donde MSE es el error cuadrático medio:

$$\text{MSE} = \sum \frac{(y_i - \hat{f}(x_i))^2}{n} \tag{2}$$

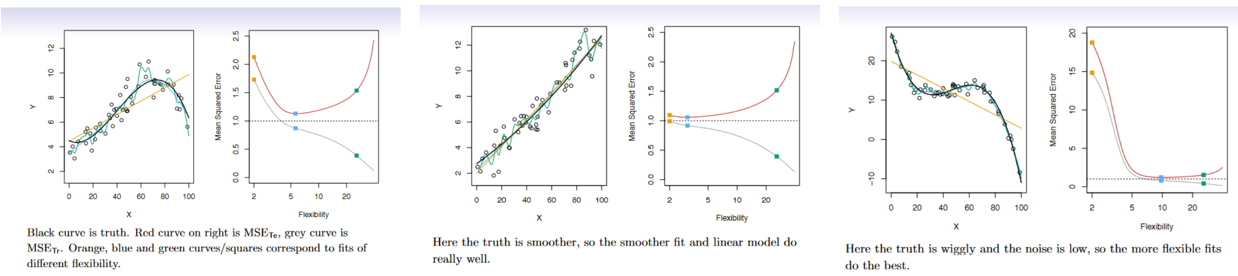


Figura 6: Distintos datos y evaluación de su MSE

En el primer caso vemos como se ha metido ruido en los datos, y vemos como el modelo ha producido overfitting. En el tercer caso, hay poco ruido y vemos como el modelo es bastante bueno.

Estos gráficos nos llevan al famoso "bias variance trade-off" (compensación entre sesgo y varianza).

El error debido al Bias de un modelo es la diferencia entre la predicción media del modelo y el valor real.

Si un modelo tiene un bias muy alto, es muy simple y no se ha ajustado bien a los datos de entrenamiento (suele ser underfitting).

La varianza de un estimador es cuánto varía los parámetros del modelo según los datos que utilicemos para el entrenamiento.

Modelo con varianza baja: cambiar los datos de entrenamiento produce cambios pequeños en la estimación. Modelo con varianza alta: pequeños cambios en el datos conlleva a grandes cambios en el output (suele ser overfitting).

2.2. Bias-Variance Trade-Off

Supongamos que hemos ajustado un modelo $\hat{f}(x)$ a algunos datos de entrenamiento Tr, y sea (x_0, y_0) una observación de prueba extraída de la población. Si el modelo verdadero es $Y = f(X) + \epsilon$ (con $f(x) = E(Y|X = x)$), entonces

$$E\left[(y_0 - \hat{f}(x_0))^2\right] = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon).$$

La expectativa promedia sobre la variabilidad de y_0 así como la variabilidad en Tr. Nótese que $\mathrm{Bias}(\hat{f}(x_0)) = E[\hat{f}(x_0)] - f(x_0)$.

Típicamente, a medida que la flexibilidad de $\hat{f}$ aumenta, su varianza aumenta y su sesgo disminuye. Así que elegir la flexibilidad basada en el error promedio de prueba implica una compensación entre sesgo y varianza.

3. Problemas de clasificación

Aquí la variable de respuesta Y es cualitativa — por ejemplo, el correo electrónico es uno de C = (spam, ham) (ham = buen correo), la clase de dígito es uno de $C = \{0, 1, \ldots, 9\}$. Nuestros objetivos son:

- Construir un clasificador $C(X)$ que asigne una etiqueta de clase de C a una futura observación X no etiquetada.
- Evaluar la incertidumbre en cada clasificación.
- Entender los roles de los diferentes predictores entre $X = (X_1, X_2, \ldots, X_p)$.

3.1. Objetivos

Sobre la base de los datos de entrenamiento nos gustaría:

- Predecir con precisión los casos de prueba no vistos.
- Entender qué entradas afectan el resultado y cómo.
- Evaluar la calidad de nuestras predicciones e inferencias.

3.2. Filosofía

- Es importante entender las ideas detrás de las diversas técnicas, para saber cómo y cuándo usarlas.
- Uno tiene que entender primero los métodos más simples, para poder comprender los más sofisticados.
- Es importante evaluar con precisión el rendimiento de un método, para saber qué tan bien o mal está funcionando [¡los métodos más simples a menudo funcionan tan bien como los más complejos!].

- Esta es un área de investigación emocionante, con aplicaciones importantes en ciencia, industria y finanzas.
- El aprendizaje estadístico es un ingrediente fundamental en la formación de un científico de datos moderno.

4. Aprendizaje no supervisado

- No hay variable de resultado, solo un conjunto de predictores (características) medidos en un conjunto de muestras.
- El objetivo es más difuso: encontrar grupos de muestras que se comporten de manera similar, encontrar características que se comporten de manera similar, encontrar combinaciones lineales de características con la mayor variación.
- Es difícil saber qué tan bien lo estás haciendo.
- Diferente del aprendizaje supervisado, pero puede ser útil como un paso de preprocesamiento para el aprendizaje supervisado.