Universidad de Santiago de Compostela

Aprendizaje estadístico

Luis Ardévol Mesa

Profesor:
Jose Almeijeiras Alonso
Manuel Mucientes Molina

Escola Técnica Superior de Enxeñaría
Master en Tecnoloxías de Análise
de Datos Masivos: Big Data

Curso 2024-2025

Contents

1 Aprendizaje estadístico

Sea una cantidad cuantitativa Y, junto con p predictores distintos, $X_1, X_2, \ldots, X_p$. Se asume una relación entre Y y $X = (X_1, X_2, \ldots, X_p)$, que se puede escribir de forma general como

$$Y = f(X) + \epsilon \tag{1.1}$$

donde f es una función multivariable desconocida de los predictores y ϵ es un término de error aleatorio, independiente de X y con media cero. En esta expresión, f representa la información sistemática que proporciona X acerca de Y.

En esencia, el aprendizaje estadístico hace referencia a una serie de métodos y aproximaciones para estimar f.

1.1 Motivos para estimar f

Algunos modelos pueden tener como objetivo la predicción y la inferencia. Por ejemplo, es el sector inmobiliario, se puede estar interesado en qué parámetros aumentan o disminuyen en mayor medida el precio de una propiedad o, conocer el valor de la misma a partir de sus características. Sin embargo, de forma general, estos problemas se tratan por separado.

1.1.1 Predicción

En ocasiones, el conjunto de entradas X está disponible fácilmente, mientras que para la salida Y ocurre lo contrario. En este contexto, como el término de error tiene media cero, se puede predecir Y a partir de

$$\hat{Y} = \hat{f}(X) \tag{1.2}$$

donde $\hat{f}$ representa la estimación de f e $\hat{Y}$ representa la predicción para Y. Generalmente, se trata a la función $\hat{f}$ como una "caja negra", en el sentido de que no interesa la forma exacta, sino como de correctas son las predicciones.

El acierto de $\hat{Y}$ como predicción para Y depende de dos cantidades: el error reducible y el irreducible. De forma general, $\hat{f}$ no será una estimación perfecta de f, y esta inexactitud introducirá cierto error. Este error es reducible porque, potencialmente, se puede mejorar la exactitud de $\hat{f}$ mediante el uso de técnicas de aprendizaje estadístico más adecuadas al problema. Sin embargo, incluso dando una estimación perfecta de f, es decir, $\hat{Y} = f(X)$, la predicción aún tendría error, ya que Y también es función de ϵ que, por definición, es impredecible usando X. Por tanto, la variabilidad asociada a ϵ también afecta a la exactitud del modelo. Este es el conocido como error irreducible, ya que, por muy buena que sea la predicción de f; no se puede reducir.

Se puede demostrar que el error irreducible es mayor que cero. La cantidad ϵ puede contener variables que no se han medido y son útiles para predecir Y; al no medirlas, f no las puede usar para predecir. Además, ϵ puede contener variaciones no mensurables; por ejemplo,

el riesgo de una reacción adversa en un paciente puede variar dependiendo del día.

Sea una estimación $\hat{f}$ y un conjunto de predictores X que conducen a una predicción $\hat{Y} = \hat{f}(X)$. Sean $\hat{f}$ y X fijos entonces, se puede comprobar que

$$E(Y - \hat{Y})^2 = E[f(X) + \epsilon - \hat{f}(X)]^2 = \underbrace{[f(X) - \hat{f}(X)]^2}_{\text{reducible}} + \underbrace{\text{Var}(\epsilon)}_{\text{irreducible}} \tag{1.3}$$

donde $E(Y - \hat{Y})^2$ representa la media, o valor esperado, de la diferencia al cuadrado entre el valor real y el predicho, y $\text{Var}(\epsilon)$ es la varianza asociada a al término de error ϵ.

El error irreducible va a proporcionar siempre una cota superior (generalmente desconocida) en la exactitud de la predicción de Y.

1.1.2 Inferencia

En ocasiones, se tiene interés en conocer cómo afectan los cambios de los predictores $X_1, \ldots, X_p$ al valor de Y. De este modo, el objetivo es estimar f para conocer la relación entre X e Y, es decir, cómo cambia Y en función de cada uno de los predictores. Ahora el problema no es necesariamente hacer predicciones de Y, y $\hat{f}$ no puede ser tratada como una "caja negra". En este contexto, interesa contestar preguntas como las siguientes:

- ¿Qué predictores están asociados con la respuesta? Generalmente, solo una fracción pequeña de los predictores disponibles están asociados de forma sustancial con Y.
- ¿Cuál es la relación de la respuesta con cada predictor?
- ¿Se puede considerar la relación entre cada predictor e Y lineal, o es más compleja? Históricamente, muchos métodos para estimar f consideran una forma lineal. En algunos casos, esto es razonable, pero en otros no basta para representar la relación entre las variables.

1.2 Estimación de f

Sea un conjunto de n observaciones, denotadas como conjunto de entrenamiento (*training data*), que se usarán para que el modelo aprenda a estimar f. Sea x_{ij} el valor del predictor j-ésimo para la i-ésima observación. Entonces, el conuunto de entrenamiento tendrá la forma $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, donde $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$.

Se busca encontrar una función $\hat{f}$ tal que $Y \approx \hat{f}(X)$ para una observación (X, Y). De forma general, los métodos de aprendizaje estadístico se pueden distinguir en dos grupos: paramétricos y no paramétricos.

1.2.1 Métodos paramétricos

Estos métodos están basados en un modelo de dos fases:

- Primero, se asume la forma funcional de f. Por ejemplo, una hipótesis simple seria que f es lineal en X

  $$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \tag{1.4}$$

  Asumiendo que la relación es lineal, se reduce el problema de estimar una función p-dimensional arbitraria a la estimación de $p + 1$ coeficientes β_μ, con $\mu = 0, 1, \ldots, p$.

- Tras la elección del modelo, se necesita un proceso que use el conjunto de entrenamiento para ajustar o entrenar el modelo. En el caso del modelo lineal, se busca encontrar los valores de los parámetros β_μ tales que

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \tag{1.5}$$

  La aproximación más común para este caso es el método de mínimos cuadrados.

Esta aproximación en dos pasos se refiere como paramétrica, al reducir el problema de estimar f a la estimación de un conjunto de coeficientes o parámetros. El problema de estos métodos es que la función elegida puede diferir en gran medida de la real, llevando a predicciones pobres. Se puede ser más flexible, ajustando a varias formas funcionales, pero esto puede conducir a un problema de *overfitting*.

1.2.2 Métodos no paramétricos

Estos métodos no hacen hipótesis explícitas sobre la forma funcional de f. Simplemente buscan una estimación de f que se aproxime lo más posible a los datos del conjunto de entrenamiento, sin ser ni muy flexible ni muy rígido. Estos tienen la ventaja de que pueden encontrar una mayor variedad de formas de f. Sin embargo, al no reducir el problema de obtener f, necesitan un gran número de observaciones para obtener una buena estimación de f.

1.2.3 Exactitud vs interpretabilidad

En inferencia, los modelos más retrictivos resultan mucho más interpretables que los flexibles, ya que dan una relación más clara entre cada predictor y la salida Y. En modelos de predicción, por el contrario, muchas veces no interesa la interpretabilidad sino la exactitud del resultado. En estos casos, se puede optar por modelos más flexibles (teniendo cuidado de no caer en problemas de *overfitting*). De forma general, a mayor flexibilidad del modelo, menor interpretabilidad.

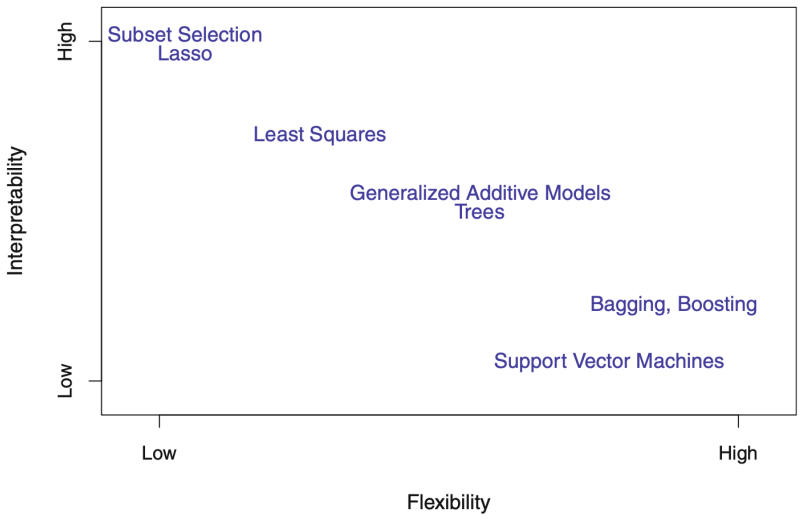

FIGURE I: Representación de la flexibilidad vs interpretabilidad de distintos métodos.

1.2.4 Aprendizaje supervisado y no supervisado

La gran mayoría de los problemas de aprendizaje estadístico entran un una de estas dos categorías: supervisado o no supervisado. Todo lo visto hasta ahora se trata de aprendizaje supervisado: para cada observación de un predictor x_i, $i = 1, \ldots, n$, hay una salida asociada y_i.

Por el contrario, el aprendizaje no supervisado describe la situación en la que para cada observación se tiene un vector de medidas x_i, pero no una salida asociada y_i. En este contexto, se busca comprender las relaciones entre las variables o entre las observaciones. Un ejemplo en el *clustering*.El objetivo de este método es establecer, en la base $x_1, \ldots, x_n$, si una observación cae dentro de grupos relativamente distintos; se limita a buscar agrupaciones de datos que distingan un subconjunto de otro. Esto se puede ver en la figura II.

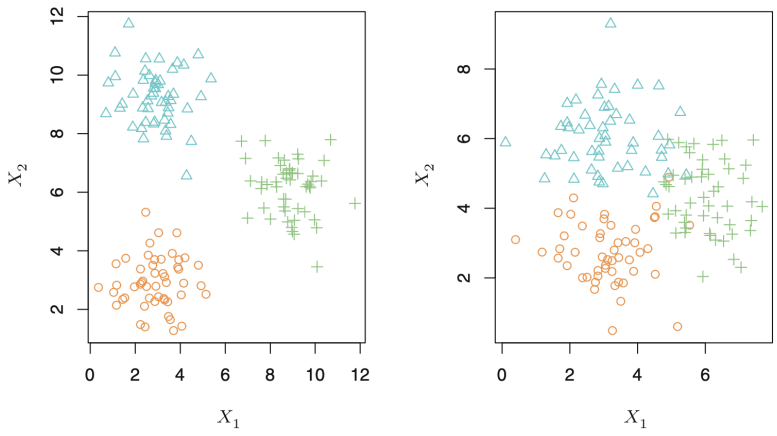

FIGURE II: *Clusters* de datos en un conjunto de observaciones. A la izquierda se observan tres grupos bien distinguidos, mientras que a la derecha la separación es más difusa.

También existen metodos semisupervisados, donde se tiene la salida correspondiente a una fracción de las observaciones y no al total. En el caso de querer considerar las n observaciones, tengan o no salida correspondiente, habría que usar este tipo de métodos.

1.2.5 Problemas de regresión y clasificación

Las variables puedes caracterizarse como cuantitativas o cualitativas (categóricas). Generalmente, los problemas que involucran respuestas cuantitativas son de regresión, mientras que los que involucran respuestas cualitativas son de clasificación (sin atender a la naturaleza de los predictores). Sin embargo, algunos métodos estadísticos, como K-vecinos más próximos, pueden usarse para mabos tipos de variables.

1.3 Exactitud del modelo

1.3.1 Calidad del ajuste

Para evaluar el rendimiento de un método de aprendizaje estadístico en un conjunto de datos determinado, es necesario cuantificar cómo de próximas son las respuestas predichas para una observación de las reales para esa misma observación. En el contexto de regresión, la medida

comúnmente más usada es el error cuadrático media (MSE),

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 \tag{1.6}$$

donde $\hat{f}(x_i)$ es la predicción de $\hat{f}$ para la observación i-ésima. El MSE será menor cuanto más cerca se encuentre la respuesta predicha de la real. Se calcula con el conjunto de datos de entrenamiento, y se denota estríctamente como MSE de entrenamiento. En la práctica, no hay mucho interés en este MSE, sino en el de *test*, es decir, cómo de próxima es la respuesta predicha a la real, para una observación nunca vista, con la que no ha entrenado. Se elige un método que minimice el MSE de *test*, ya que un MSE de entrenamiento bajo no garantiza un MSE de *test* bajo. Para un número grande observaciones de *test*, se puede calcular la error cuadrático medio de la predicción para las oservaciones de *test*, (x_0, y_0),

$$\text{Ave}(y_0 - \hat{f}(x_0))^2 \tag{1.7}$$

Cuando un modelo tiene un MSE de entrenamiento bajo, pero uno de *test* alto, se tiene un problema de *overfitting*. Sin embargo, el cálculo del MSE de *test* puede ser complicado debido a la ausencia, en muchos casos, de un conjunto de *test*.

1.3.2 Compensación entre *bias* y varianza

Se puede demostrar que el MSE de *test* esperado, para un valor x_0 dado, puede descomponerse en la suma de tres cantidades fundamentales: la varianza de $\hat{f}(x_0)$, el *bias* al cuadrado de $\hat{f}(x_0)$ y la varianza del término de error ϵ. Esto es

$$MSE \equiv E(y_0 - \hat{f}(x_0))^2 = \text{Var}(\hat{f}(x_0)) + [\text{Bias}(\hat{f}(x_0))]^2 + \text{Var}(\epsilon) \tag{1.8}$$

Aquí, el MSE de *test* esperado es el promedio de numerosas estimaciones de f usando un gran número de conjuntos de entrenamiento, comprobados todos sobre x_0. El primer término se refiere a la cantidad en la que cambiaría $\hat{f}$ si se estimara usadno un conjunto diferente de entrenamiento. El término de *bias* hace referencia al error introducido por aproximar el problema real por uno mucho más simple, y el último término se corresponde con el error irreducible.

La ecuación (1.8) muestra que, para minimizar el error de *test* esperado, se necesita seleccionar un método de aprendizaje estadístico que consiga baja varianza y bajo *bias*. La varianza es, por definición, no negativa, así como el cuadrado del *bias*. De este modo, el MSE esperado de *test* nunca puede ser menor que $\text{Var}(\epsilon)$.

Generalmente, cuánto más flexible sea el método usado, mayor será la varianza y menor el *bias*. Así, el cambio relativo entre ambas determina si el MSE de *test* aumenta o disminuye. Esta relación se puede ver en la figura III.

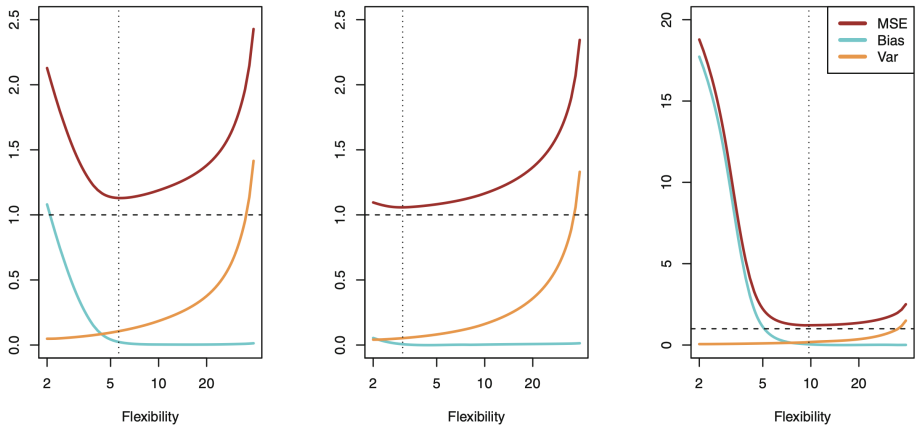

FIGURE III: *Bias* al cuadrado (azul), varianza (naranja), Var(ϵ) (línea discontinua) y MSE de *test* (rojo) para tres conjuntos de datos distintos. Las líneas punteadas verticales indican el nivel de flexibilidad correspondiente al menor MSE de *test*.

2 Regresión lineal

Ejemplo 2.0.1. *Supongamos un conjunto de datos de publicidad con las ventas de un producto como función del presupuesto destinado a publicidad en radio, TV y periódico. Algunas preguntas importantes que deberíamos contestar serían, por ejemplo:*

- *¿Existe alguna relación entre el presupuesto en publicidad y las ventas?*
- *¿Cómo de relacionados están el presupuesto en publicidad y las ventas?*
- *¿Que media contribuye más a las ventas?*
- *¿Cómo de exacto podemos estimar el efecto de cada medio en las ventas?*
- *¿Cómo de exacto podemos predecir futuras ventas?*
- *¿La relación es lineal?*

2.1 Regresión lineal simple

Esta es una aproximación para predecir una respuesta cuantitativa Y en base a un único predictor X, asumiendo una relación lineal entre ambos

$$Y \approx \beta_0 + \beta_1 X \tag{2.1}$$

donde β_0 y β_1 son coeficientes desconocidos que representan la ordenada en el origen y la pendiente de la recta, y que se denotarán como coeficientes o parámetros del modelo. Una vez usado el conjunto de entrenamiento para estimar $\hat{\beta}_0$ y $\hat{\beta}_1$, se pueden hacer predicciones de la forma

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{2.2}$$

donde $\hat{y}$ es la predicción de Y basada en $X = x$.

2.1.1 Estimación de los coeficientes

En la práctica, β_0 y β_1 son desconocidos. Sea un conjunto de n observaciones $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, donde cada pareja consiste de una medida de X y otra de Y. Para obtener los coeficientes estimados $\hat{\beta}_0$ y $\hat{\beta}_1$, se pueden usar varios métodos. Primero veremos el método de minimización de mínimos cuadrados.

Sea $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ la predicción para Y basada en el valor i-ésimo de X. Entonces, $e_i = y_i - \hat{y}_i$ representa el residuo i-ésimo, es decir, la diferencia entre el valor observado y el predicho.Se define la suma de residuos al cuadrado (RSS) como

$$RSS = e_1^2 + e_2^2 + \cdots + e_n^2 = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \tag{2.3}$$

La aproximación de mínimos cuadrados toma los $\hat{beta}$ que minimizan el RSS, que tienen la siguiente forma

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{2.4}$$

siendo $\bar{x}$ y $\bar{y}$ las medias de x e y, respectivamente.

2.1.2 Exactitud de la estimación de los coeficientes

El problema inicial es estimar la relación f entre las respuestas Y y un conjunto de predictores X. Si se asume que esta relación es lineal, se puede reescribir el problema de la siguiente forma

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{2.5}$$

Generalmente, se asume que el término de error es independiente de X. La expresión (2.5) define la recta de regresión poblacional, que resulta la mejor aproximación lineal de la relación real entre X e Y. Los coeficientes $\hat{\beta}$ carcaterizan la recta de mínimos cuadrados (2.2). Esta es la recta a la que se tiene acceso generalmente, no la poblacional.

La analogía entre regresión lineal y la estimación de la media de una variable aleatoria se fndamenta en el concepto de sesgo o *bias*. Si se usa la media de la muestra $\hat{\mu}$ para estimar μ, esta estimación no presenta sesgo (*unbiased*), en el sentido de que, en media, se espera que $\hat{\mu}$ sea igual a μ. Esto quiere decir que para un conjunto de observaciones $y_1, \ldots, y_n$, $\hat{\mu}$ puede estar sobreestimando μ, mientras que para otro conjunto de observaciones, se puede estar subestimando. Si se pudiera promediar un gran número de conjuntos de observaciones, $\hat{\mu} = \mu$. Así, un estimador no sesgado no sobreestime ni subestima sistemáticamente los parámetros. Este mismo concepto aplica a la estimación de β_0 y β_1: si se pudiera promediar un gran número de conjuntos de datos, entonces el resultado sería $\hat{\beta} = \beta$.

Continuando con esta analogía, se busca conocer cómo de precisa es la estimación $\hat{\mu}$ de μ. En general, la desviación de una única estiamción $\hat{\mu}$ respecto a μ vendrá dada por el error estandar[1],

$$\mathrm{Var}(\hat{\mu}) = \mathrm{SE}(\hat{\mu})^2 = \frac{\sigma^2}{n} \tag{2.6}$$

donde σ es la desviación estándar de cada realización y_i de Y. El error estándar asociado a los coeficientes de mínimos cuadrados, es decir, cómo de cercanos son $\hat{\beta}_0$ y $\hat{\beta}_1$ a β_0 y β_1,

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right], \qquad \mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2.7}$$

donde $\sigma^2 = \mathrm{Var}(\epsilon)$. Para que estas fórmulas sean válidas, se debe asumir que los errores ϵ_i de cada observación no están relacionados con la varianza común σ^2. Aunque en algunos casos no se cumpla, resulta una buena aproximación en general. En muchas ocasiones, σ^2 es desconocida, pero se puede estimar a partir de los datos. Esta estimación se conoce como error estándar residual

$$RSE = \sqrt{\frac{\mathrm{RSS}}{n-2}} \tag{2.8}$$

Los errores estándar se pueden usar para calcular intervalos de confianza. Un intervalo de confianza al 95% se define como un rango de valores tal que, con un 95% de probabilidad, el valor real y desconocido del parámetro estará ahí contenido. El rango queda definido por una cota superior e inferior que se calcula a través de la muestra. En el caso de regresión lineal, el intervalo de confianza al 95% para β_0 y β_1 es similar

$$\hat{\beta}_i \pm 2 \cdot \mathrm{SE}(\hat{\beta}_i) \tag{2.9}$$

[1] el error estándar es la desviación estándar de la distribución muestral de un estadístico muestral.

Los errores estándar también pueden ser usados para realizar contrastes de hipótesis acerca de los coeficientes. El contraste más común implica conctrastar la hipótesis nula

$$H_0 : \text{No existe relación entre } X \text{ e } Y \tag{2.10}$$

contra la hipótesis alternativa

$$H_a : \textit{Existe alguna relación entre} X \text{ e } Y \tag{2.11}$$

Matemáticamente, esto es

$$H_0 : \beta_1 = 0 \tag{2.12}$$
$$H_a : \beta_1 \neq 0 \tag{2.13}$$

ya que, si $\beta_1 = 0$, el modelo se reduce a $Y = \beta_0 + \epsilon$, por lo que no habría relación entre las variables. Para probar la hipótesis nula, senecesita determinar si $\hat{\beta}_1$ es suficientemente grande como para tener la confianza de que β_1 es distinto de cero. El tamaño dependerá de la precisión de $\hat{\beta}_1$, es decir, dependerá de $\text{SE}(\hat{\beta}_1)$. Si este valor es pequeño, entonces valores pequeños de $\hat{\beta}_1$ probarían que $\beta_1 \neq 0$ Si el valor es grande, se tiene el caso contrario. En la práctica, se hace un *t-test* dado por

$$t = \frac{\hat{\beta}_1 - 0}{\text{SE}(\hat{\beta}_1)} \tag{2.14}$$

que mide el número de desviaciones estándar que $\hat{\beta}_1$ se aleja de cero. Si no hay relación entre X e Y, se espera tener una t-distribución con $n-2$ grados de libertad. La t-distribución tiene forma de campana y para valores de n mayores a de 30, aproximadamente, es similar a una distribución normal. Consecuentemente, es sencillo calcular la probabilidad de observar un valor igual a $|t|$ o mayor, asumiendo $\beta_1 = 0$; esta probabilidad se conoce como p-valor. De forma poco técnica, el p-valor se interpreta de la sigueinte forma: un p-valor pequeño indica que es poco probable observar dicha asociación entre el predictor y la respuesta. Típicamente se fija una cota, la significancia; si el p-valor es menor que esta cota, se rechaza la hipótesis nula.

2.1.3 Exactitud del modelo

La calidad de un modelo de regresión lineal se suele cuantificar mediante dos cantidades, el error estándar residual (RSE) y el coeficiente R^2.

Error estándar residual

El RSE es una medida de la desviación estándar del término de error ϵ en el modelo. Se puede ver como la cantidad media que se desviará la respuesta de la recta real de regresión, y se calcula como

$$RSE = \sqrt{\frac{RSS}{n-2}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2.15}$$

El RSE se considera una medida de la "falta de ajuste" del modelo a los datos. Si $\hat{y}_i \approx y_i$, entonces el RSE será pequeño, por lo que el modelo ajustará bien los datos.

Estadística R^2

El RSE proporciona una medida absoluta de la falta de ajuste del modelo a los datos. Sin embargo, al estar medido en las unidades de Y, no siempre queda claro qué constituye un buen RSE. La estadística R^2 da una medida alternativa acerca del ajuste. Tiene la forma de una proporción, por lo que toma valores entre 0 y 1, y resulta independiente de la escala de Y.

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{RSS}} = 1 - \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2.16}$$

donde TSS es la suma total de cuadrados, que mide la varianza total en la respuesta Y, y se puede ver como la cantidad de variabilidad inherente en la respuesta antes de hacer la regresión. Por el contrario, el RSS mide la cantidad de variabilidad que queda sin explicación tras hacer la regresión. Así, TSS − RSS mide la cantidad de variabilidad en la respuesta que es explicada (o eliminada) al hacer la regresión, y R^2 mide la proporción de variabilidad en Y que puede explicarse usando X. Si R^2 es cercano a 1, significa que una gran proporción de la variabilidad en la respuesta fue explicada por la regresión. Un R^2 cercano a 0 puede ocurrir porque el modelo esté mal, porque el error inherente σ^2 es grande, o ambas.

R^2 es una medida de la relación lineal entre X e Y. La correlación, definida como

$$\text{Cor}(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2.17}$$

también es una medida de la relación lineal entre X e Y. Esto sugiere que se podría usar $r = \text{Cor}(X, Y)$ en lugar de R^2 como estimador del ajuste del modelo lineal. Se puede demostrar que, en el caso de un modelo de regresión lineal simple, $R^2 = r^2$.

2.2 Regresión multilineal

En general, se tienen p predictores distintos. El modelo de regresión multilineal toma la forma

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon \tag{2.18}$$

donde X_j representa el predictor j-ésimo y β_j cuantifica la relación entre esa variable y la respuesta. Se puede interpretar β_j como el efecto medio sobre Y al incrementar X_j en una unidad, manteniendo el resto de predictores fijos. Ahora, el conjunto de entrenamiento tendrá la siguiente forma

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} \tag{2.19}$$

2.2.1 Estimación de los coeficientes

Como en el caso de la regresión simple, los coeficientes $\beta_0, \ldots, \beta_p$ son desconocidos, y sus estimaciones, $\hat{\beta}_0, \ldots, \hat{\beta}_p$, permiten hacer predicciones usando

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p \tag{2.20}$$

Los parámetros se estiman por el mismo método de mínimos cuadrados visto anteriormente. Se eligen $\beta_0, \beta_1, \ldots, \beta_p$ de modo que se minimize la suma de los residuos al cuadrado

$$\text{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2 = ||\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}||^2 \tag{2.21}$$

donde

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \quad \hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{pmatrix} \tag{2.22}$$

En este caso, es más complicado dar el cálculo de estos coeficientes, pero se puede demostrar que los valores de $\hat{\boldsymbol{\beta}}$ que minimizan el RSS tienen la forma

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{2.23}$$

Hay que tener en cuenta que un modelo lineal simple puede mostrar una relación entre un predictor y la respuesta, y el modelo multilineal afirmar lo contrario. Por ejemplo, en el caso de la publicidad, existe una en un modelo lineal simple de regresión, existe una relación entre las ventas y el gasto en periódico, pero en el modelo de regresión multilineal se ve que no hay relación directa. Estos casos ocurren porque uno es el sustituto del otro, es decir, "se lleva" el crédito. En el ejemplo de la publicidad, el periódico se llevaría el crédito del efecto de la radio.

2.2.2 Algunas preguntas importantes

Al hacer una regresión multilineal, normalmente se busca contestar las siguientes cuestiones

Relación entre la respuesta y los preditores

De forma análoga al caso de la regresión simple, la relación entre los predictores y la respuesta se puede comprobar mediante un contraste de hipótesis,

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \tag{2.24}$$

$$H_a : textalmenosun\beta_j \text{ es no nulo} \tag{2.25}$$

El *test* se realiza calculando una *F-statistic*,

$$F = \frac{(\text{TSS} - \text{RSS})/p}{\text{RSS}/(n-p-1)} \tag{2.26}$$

Si la hipótesis de modelo lineal es correcta, se puede demostrar que

$$E\{\text{TRSS}/(n-p-1)\} = \sigma^2 \tag{2.27}$$

y, si H_0 es cierto,

$$E\{(\text{TSS} - \text{RSS})/p\} = \sigma^2 \tag{2.28}$$

Así, cuando no existe relación entre los predictores y la respuesta, se espera un valor del *F-statistic* próximo a 1. Por el contrario, si H_a es cierto, se espera un valor mayor que 1 ya que

$$E\{\text{TRSS}/(n-p-1)\} > \sigma^2 \tag{2.29}$$

Para un p-valor pequeño, se puede rechazar la hipótesis nula. La *F-statistic* se puede realizar sobre un subconjunto de q parámetros, con $q < p$. En este caso, la ecuación (2.26) tomaría la forma

$$F = \frac{(\text{RSS}_0 - \text{RSS})/q}{\text{RSS}/(n-p-1)} \tag{2.30}$$

donde RSS_0 es la suma de residuos al cuadrado de ese modelo. Capitulo 6 para cuando $p > n$

Variables relevantes

Tras realizar una *F-statistic* y concluir que existe relación entre los predictores y la respuesta, el siguiente paso es conocer qué predictores están involucrados. Se puede atender a los '-valores individuales, pero para valores de p grandes, esto dará falsos positivos. Capitulo6

Ajuste del modelo

Dos de las medidas más comunes para cuantificar el ajuste del modelo son el RSE y R^2.

En regresión múltiple, R^2 coincide con $\text{Cor}(Y, \hat{Y})^2$, la correlación entre la respuesta real y la predicha por el modelo lineal al cuadrado. Se puede demostrar que un modelo bien ajusta es aquel que maximiza esta correlación entre todos los posibles modelos lineales.

Un valor de R^2 próximo a 1 indica que el modelo explica una gran proporción de la varianza en la variable de respuesta. Además, este coeficiente siempre crecerá cuando se añadan más variables al modelo. Un aumento muy sútil puede sugerir que la variable recién añadida se puede obviar en el modelo.

De forma general, el RSE se define como

$$\text{RSE} = \sqrt{\frac{\text{RSS}}{n-p-1}} \tag{2.31}$$

y en el caso de regresión simple se simplifica a (2.15). Así, los modelos con más variables pueden tener un RSE alto si la diminución en el RSS es pequeña en relación al incremento en p.

Predicciones

Tras ajustar el modelo de regresión multilineal, obtener las predicciones aplicando (2.20) es directo. Sin embargo, esta predicción tiene tres tipos de incertidumbres asociadas:

i. Los coeficientes $\hat{\beta}$ son estimaciones de β, es decir, el plano de mínimos cuadrados

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p \tag{2.32}$$

es solo una estimación del plano de regresión poblacional real

$$f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \tag{2.33}$$

La falta de exactitud en la estimación de los coeficientes está relacionada con el error reducilbe. Se pueden calcular los intervalos de confianza para determinar cómo de próximo es $\hat{Y}$ a $f(X)$.

ii. En general, asumir un modelo lineal es una aproximación, lo que resulta en un *bias* para el modelo. Esto es una fuente de error potencialmente reducible.

iii. Incluso conociendo $f(X)$, siempre existirá el error aleatorio ϵ del modelo, es decir, el error irreducible. Para ver como diferirán Y e $\hat{Y}$, se usan intervalos de predicción. Estos intervalos son siempre más amplios que los de confianza, ya que incorporan el error en la estimación de $f(X)$ y la incertidumbre en cada punto dada por el error irreducible.

2.2.3 Problemas potenciales

Los problemas más comunes al asumir un modelo lineal son los siguientes

- No linealidad en la relación respuesta-predictor.
- Correlación de los términos de error.
- Varianza de los términos de error no constante.
- Datos atípicos (*outliers*).
- Puntos de gran influencia.
- Colinelidad (multicolinealidad).

2.3 Grandes conjuntos de variable correlacionadas

Uno de los problemas más comunes en regresión lineal es la colinealidad, es decir, la alta correlación entre predictores. La colinealidad puede ser un problema en regresión lineal porque puede aumentar la varianza de los coeficientes estimados, lo que puede llevar a una gran sensibilidad de los coeficientes a pequeños cambios en el modelo. Además, puede ser difícil interpretar los coeficientes cuando los predictores están altamente correlacionados. Para afrontar este problema, existe varias técnicas:

- Selección de subconjuntos. Métodos para seleccionar un subconjunto de predictores. Luego se ajusta un modelo usando mínimos cuadrados en el conjunto reducido de variables.
- Regularización. Esta aproximacion supone ajustar un modelo que contenga todos los predictores, pero que penalice o regularice los coeficientes estimados (hacia cero).
- Reducción de dimensión. Métodos para proyector los p predictores sobre un subespacio de menor dimensión. Luego, las proyecciones se usan como predictores para ajustar un modelo de regresión lineal.

En las próximas secciones se describirán los métodos de regularización y reducción de dimensión. La selección de subconjuntos se tratará más adelante.

2.4 Métodos de reducción

Como alternativa a los métodos de selección de subconjuntos que se verán más adelante, se puede ajustar un modelo que contenga los p predictoes usando una técnica que penalice o regularice los coeficientes estimados o, equivalentemente, que encoja las estimaciones hacia cero. Esto puede reducir la varianza de los coeficientes estimados significativamente. A continuación,

se describen dos métodos más comunes para encoger los coeficientes hacia cero: la regresión de Ridge y la regresión Lasso. El problema que resuelve la estimación por mínimos cuadrados es

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \quad \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 \tag{2.34}$$

2.4.1 Regresión de Ridge

La regresión de Ridge es similar a los mínimos cuadrados, excepto que los coeficientes se estiman minimizando una cantidad diferente. En particular, los coeficientes estimados por la regresión Ridge, $\hat{\beta}^R$, son los valores que minimizan, siguiendo el método de los multiplicadores de Lagrange,

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = \text{RSS} + \lambda\sum_{j=1}^{p}\beta_j^2 \tag{2.35}$$

donde $\lambda \geq 0$ es un parámetro de ajuste que debe determinarse. Esto puede escribirse como

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \quad \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 \tag{2.36}$$

$$\text{sujeto a} \quad \sum_{j=1}^{p}\beta_j^2 \leq s \tag{2.37}$$

Como con el método de mínimos cuadrados, la regresión Ridge busca estimaciones de coeficientes que ajusten bien los datos, minimizando la RSS. Sin embargo, la regresión Ridge también añade un segundo término a la RSS que penaliza la magnitud de los coeficientes estimados. El parámetro λ controla la cantidad de penalización: cuanto mayor sea λ, mayor será la penalización (valores más encogidos). Así, el parámetro de ajuste controla el impacto relativo de estos dos términos en la estimación de los coeficientes de regresión. Si $\lambda = 0$, entonces la penalización es nula, y la regresión Ridge se convierte en mínimos cuadrados. Si $\lambda \to \infty$, entonces la penalización crece y los coeficientes estimados por Ridge se aproximan a cero. A diferencia de mínimos cuadrados, que genera un único conjunto de coeficientes estimados, la regresión Ridge producirá diferentes conjuntos de estimaciones $\hat{\beta}^R$ para cada valor de λ. Por tanto, elegir un valor adecuado de λ es crítico.

Nótese que no se encoge β_0, ya que simplemente es una medida del valor medio de la respuesta cuando $x_{i1} = x_{i2} = \cdots = x_{ip} = 0$. Si los predictores han sido centrados para tener media cero antes de realizar la regresión Ridge, entonces β_0 tendrá la forma

$$\hat{\beta}_0^R = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{2.38}$$

Las estimaciones estándar de los coeficientes de mínimos cuadrados ya discutidas son equivariantes a escala: multiplicar X_j por una constante c simplemente lleva a una escala de las estimaciones de los coeficientes de mínimos cuadrados por un factor de $\frac{1}{c}$. En otras palabras, independientemente de cómo se escale el predictor X_j, $\hat{X}_j\beta_j$ permanecerá igual. En contraste, las estimaciones de los coeficientes de regresión Ridge pueden cambiar sustancialmente al multiplicar un predictor dado por una constante; no son equivariantes a escala. Por tanto, se

deben estandarizar las entradas antes de resolver el problema de optimización con la ecuación

$$\tilde{x}_{ij} = \frac{x_i j}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}} \tag{2.39}$$

Los coeficientes estimados por regresión Ridge tienen el menor RSS de todos los puntos que se encuentran en el hiperboloide cerrado definido por $\sum_{i=1}^{p} \beta_j^2 y \geq s$. En el caso de tener dos predictores, $p = 2$, esto no es más que los puntos que estén dentro y sobre la superficie del círculo de la figura IV.

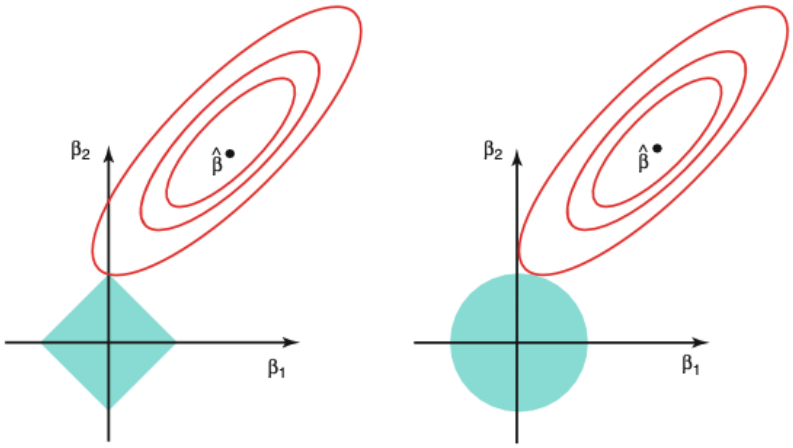

FIGURE IV: Contornos de las funciones de error y ligadura para la regresión Lasso (izquierda) y Ridge (derecha). Las áreas de color azul son las regiones de ligadura $|\beta_1| + |\beta_2| \leq s$ y $\beta_1^2 + \beta_2^2 \leq s$, mientras que las elipses rojas son los contornos del RSS.

Mejora de Ridge sobre mínimos cuadrados

La ventaja de la regresión Ridge sobre los mínimos cuadrados se basa en el compromiso entre sesgo y varianza. A medida que λ aumenta, la flexibilidad del ajuste de la regresión Ridge disminuye, lo que lleva a una disminución de la varianza pero a un aumento del sesgo. Esto se ilustra en el panel izquierdo de la figura V. La curva verde en el panel izquierdo de la Figura 6.5 muestra la varianza de las predicciones de la regresión Ridge como una función de λ. En las estimaciones de los coeficientes de mínimos cuadrados, que corresponden a la regresión Ridge con $\lambda = 0$, la varianza es alta pero no hay sesgo. Sin embargo, a medida que λ aumenta, la contracción de las estimaciones de los coeficientes Ridge conduce a una reducción sustancial de la varianza de las predicciones, a costa de un leve aumento en el sesgo. Recuerda que el error cuadrático medio de prueba (MSE), graficado en púrpura, es una función de la varianza más el sesgo al cuadrado. Para valores de λ de hasta aproximadamente 10, la varianza disminuye rápidamente, con un aumento muy leve en el sesgo, graficado en negro. En consecuencia, el error cuadrático medio (MSE) disminuye considerablemente a medida que λ aumenta de 0 a 10. Más allá de este punto, la disminución en la varianza debido al aumento de λ se ralentiza, y la contracción en los coeficientes causa que se subestimen significativamente, resultando en un gran aumento en el sesgo. El MSE mínimo se alcanza aproximadamente en $\lambda = 30$. Curiosamente, debido a su alta varianza, el MSE asociado con el ajuste de mínimos cuadrados, cuando $\lambda = 0$, es casi tan alto como el del modelo nulo, en el cual todas las estimaciones de

los coeficientes son cero, cuando $\lambda \to \infty$. Sin embargo, para un valor intermedio de λ, el MSE es considerablemente menor.

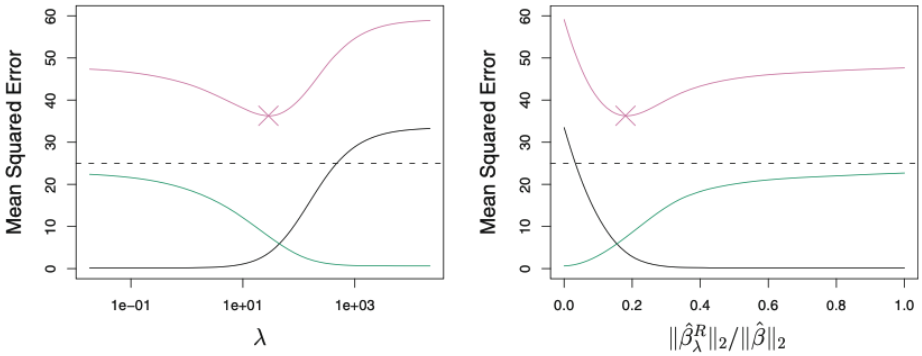

FIGURE V: *Bias* al cuadrado (negro), varianza (verde) y MSE de *test* (morado) para predicciones de regresión Ridge en un conjunto de datos simulado, como función de λ y $||\hat{\beta}^R_\lambda||_2/||\hat{\beta}||_2$. La linea discontinua horizontal indica el mínimo MSE posible. Las cruces moradas indican los modelos de regresión Risge para los cuales el MSE es el menor.

El panel derecho de la figura V muestra las mismas curvas que el panel izquierdo, esta vez graficadas contra la norma ℓ_2 de las estimaciones de los coeficientes de la regresión Ridge dividida por la norma ℓ_2 de las estimaciones de mínimos cuadrados. Ahora, a medida que se va de izquierda a derecha, los ajustes se vuelven más flexibles, por lo que el sesgo disminuye y la varianza aumenta.

En general, en situaciones donde la relación entre la respuesta y los predictores es casi lineal, las estimaciones por mínimos cuadrados tendrán un sesgo bajo pero pueden tener una alta varianza. Esto significa que un pequeño cambio en los datos de entrenamiento puede causar un gran cambio en las estimaciones de los coeficientes por mínimos cuadrados. En particular, cuando el número de variables p es casi tan grande como el número de observaciones n, como en el ejemplo de la figura V, las estimaciones por mínimos cuadrados serán extremadamente variables. Y si $p > n$, entonces las estimaciones por mínimos cuadrados ni siquiera tienen una solución única, mientras que la regresión Ridge aún puede funcionar bien al intercambiar un pequeño aumento en el sesgo por una gran disminución en la varianza. Por lo tanto, la regresión Ridge funciona mejor en situaciones donde las estimaciones por mínimos cuadrados tienen alta varianza.

Regresión Ridge también tiene una ventaja computacional sobre la elección del mejor subconjunto, que requiere buscar sobre 2^p modelos. Para un valor dado de λ, la regresión Ridge solo ajusta un único modelo. Incluso para valores moderados de p, dicha búsqueda puede ser computacionalmente inviable. En contraste, para cualquier valor fijo de λ, la regresión Ridge solo ajusta un único modelo, y el procedimiento de ajuste del modelo puede realizarse bastante rápido. De hecho, se puede demostrar que los cálculos necesarios para resolver (2.35), simultáneamente para todos los valores de λ, son casi idénticos a los de ajustar un modelo usando mínimos cuadrados.

2.4.2 Regresión Lasso

La regresión de Ridge tiene una desventaja obvia. A diferencia de la selección del mejor subconjunto, la selección hacia adelante y la selección hacia atrás, que generalmente seleccionan

modelos que involucran solo un subconjunto de las variables, la regresión Ridge incluirá todos los p predictores en el modelo final. La penalización $\lambda \sum \beta_j^2$ en (2.35) reducirá todos los coeficientes hacia cero, pero no establecerá ninguno de ellos exactamente en cero (a menos que $\lambda = \infty$. Esto puede no ser un problema para la precisión de la predicción, pero puede crear un desafío en la interpretación del modelo en contextos donde el número de variables p es bastante grande. Sin embargo, se puede querer un modelo que incluya solo los predictores más relevantes.

Lasso es una alternativa relativamente reciente a la regresión de Ridge que supera esta desventaja. Los coeficientes de Lasso, $\hat{\beta}_\lambda^L$, minimizan la cantidad

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| = \text{RSS} + \lambda \sum_{j=1}^{p} |\beta_j| \tag{2.40}$$

Esto puede escribirse como

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \quad \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \tag{2.41}$$

$$\text{sujeto a} \quad \sum_{j=1}^{p} |\beta_j| \leq s \tag{2.42}$$

Comparando (2.40) con (2.35), se ve que Lasso y Ridge tienen formulaciones similares. La única diferencia es la condición de ligadura. En términos estadísticos, Lasso usa una penalización ℓ_1 en lugar de la ℓ_2. La norma ℓ_1 de un vector de coeficientes β viene dada por $||\beta||_1 = \sum |\beta_j|$.

Al igual que en la regresión Ridge, Lasso reduce las estimaciones de los coeficientes hacia cero. Sin embargo, en el caso de Lasso, la penalización ℓ_1 tiene el efecto de forzar que algunas de las estimaciones de los coeficientes sean exactamente cero cuando el parámetro de ajuste λ es suficientemente grande. Por lo tanto, al igual que la selección del mejor subconjunto, Lasso realiza una selección de variables. Como resultado, los modelos generados por Lasso son generalmente mucho más fáciles de interpretar que los producidos por la regresión Ridge. Se dice que Lasso produce modelos escasos (*sparse*) decir, modelos que involucran solo un subconjunto de las variables. Al igual que en la regresión Ridge, seleccionar un buen valor de λ para Lasso es crucial.

Propiedad de selección de variables en Lasso

¿Por qué Lasso, a diferencia de la regresión Ridge, resulta en estimaciones de coeficientes que son exactamente iguales a cero? Las formulaciones (2.37) y (2.42) pueden usarse para arrojar luz sobre el asunto. La figura IV ilustra la situación. La solución de mínimos cuadrados está marcada como $\hat{\beta}$, mientras que el diamante y círculo azul representan las restricciones de Lasso y Ridge en (2.37) y (2.42), respectivamente. Si s es suficientemente grande, entonces las regiones de restricción contendrán $\hat{\beta}$, y por lo tanto, las estimaciones de regresión Ridge y Lasso serán las mismas que las estimaciones de mínimos cuadrados (un valor tan grande de s corresponde a $\lambda = 0$ en (2.35) y (2.40)). Sin embargo, en la figura IV las estimaciones de mínimos cuadrados se encuentran fuera del diamante y del círculo, y por lo tanto, las estimaciones de mínimos cuadrados no son las mismas que las estimaciones de Lasso y regresión

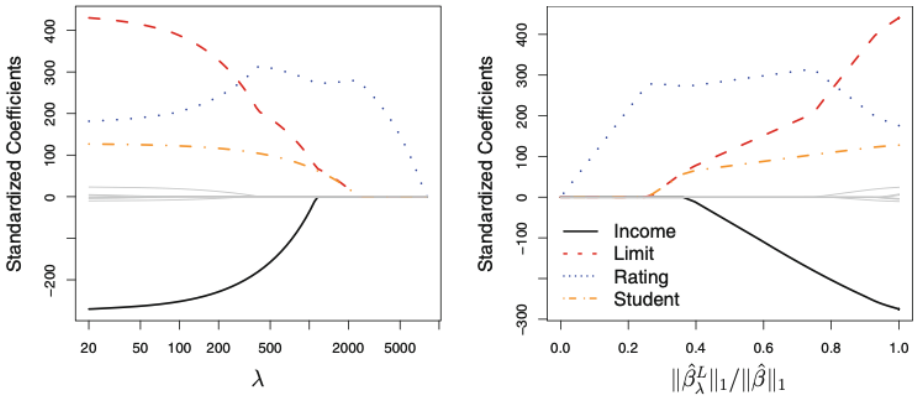

FIGURE VI: Ejemplo de coeficientes Lasso en un *dataset* como función de λ y $||\hat{\beta}^L_\lambda||_1/||\hat{\beta}||_1$. Cuando $\lambda = 0$, Lasso proporciona el ajuste de mínimos cuadrados, y cuando λ se vuelve suficientemente grande, Lasso da el modelo nulo en el que todas las estimaciones de los coeficientes son iguales a cero. Sin embargo, entre estos dos extremos, los modelos de regresión Ridge y Lasso son bastante diferentes entre sí. Moviéndose de izquierda a derecha en el panel derecho, se observa que al principio Lasso resulta en un modelo que contiene solo el predictor *rating*. Luego, *student* y *limit* entran en el modelo casi simultáneamente, seguidos poco después por *income*. Eventualmente, las variables restantes entran en el modelo. Por lo tanto, dependiendo del valor de λ, Lasso puede producir un modelo que involucre cualquier número de variables. En contraste, la regresión Ridge siempre incluirá todas las variables en el modelo, aunque la magnitud de las estimaciones de los coeficientes dependerá de λ.

Ridge.

Las elipses que están centradas alrededor de $\hat{\beta}$ representan regiones de RSS constante. En otras palabras, todos los puntos en una elipse dada comparten un valor común del RSS. A medida que las elipses se expanden alejándose de las estimaciones de los coeficientes de mínimos cuadrados, el RSS aumenta. Las ecuaciones (2.37) y (2.42) indican que las estimaciones de los coeficientes de Lasso y regresión Ridge están dadas por el primer punto en el que una elipse contacta la región de restricción. Dado que la regresión Ridge tiene una restricción circular sin puntos agudos, esta intersección generalmente no ocurrirá en un eje, y por lo tanto, las estimaciones de los coeficientes de regresión Ridge serán exclusivamente diferentes de cero. Sin embargo, la restricción de Lasso tiene esquinas en cada uno de los ejes, y por lo tanto, la elipse a menudo intersectará la región de restricción en un eje. Cuando esto ocurre, uno de los coeficientes será igual a cero. En dimensiones más altas, muchas de las estimaciones de los coeficientes pueden ser iguales a cero simultáneamente. En la figura IV, la intersección ocurre en $\beta_1 = 0$, y por lo tanto, el modelo resultante solo incluirá β_2.

En la figura IV, se considera el caso simple de $p = 2$. Cuando $p = 3$, la región de restricción para la regresión Ridge se convierte en una esfera, y la región de restricción para el lasso se convierte en un poliedro. Cuando $p > 3$, uno de los coeficientes será igual a cero. En dimensiones más altas, muchas de las estimaciones de los coeficientes pueden ser iguales a cero simultáneamente. En la figura IV, la intersección ocurre en $\beta_1 = 0$, y por lo tanto, el modelo resultante solo incluirá β_2.

En la figura IV, se considera el caso simple de $p = 2$. Cuando $p = 3$, la restricción para

la regresión Ridge se convierte en una hiperesfera, y la región de restricción para Lasso se convierte en un politopo. Sin embargo, las ideas clave representadas en la figura IV siguen siendo válidas. En particular, Lasso conduce a la selección de características cuando $p > 2$ debido a las esquinas agudas del poliedro o politopo.

Comparación entre regresión Lasso y Ridge

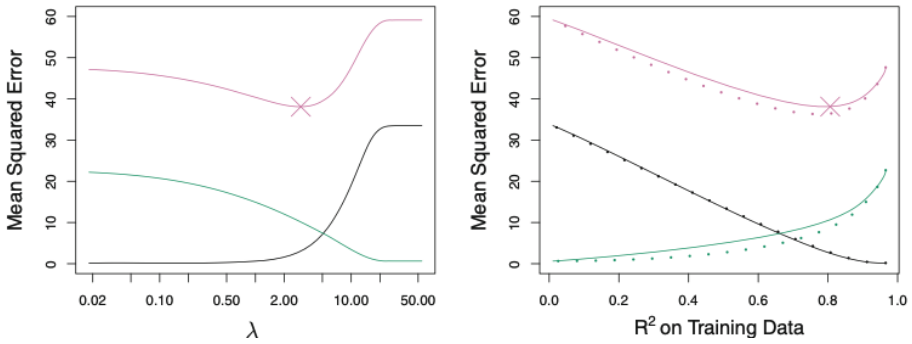

FIGURE VII: Izquierda: Gráficas del sesgo cuadrado (negro), varianza (verde) y MSE de prueba (morado) para Lasso en un conjunto de datos simulado. Los datos simulados son similares a los de la figura VII, excepto que ahora solo dos predictores se relacionan con la respuesta. Derecha: Comparación del sesgo cuadrado, varianza y MSE de prueba entre Lasso (sólido) y Ridge (discontinuo). Ambos se grafican contra su R^2 en los datos de entrenamiento, como una forma común de indexación. Las cruces en ambas gráficas indican el modelo Lasso para el cual el MSE es el más pequeño.

Es claro que Lasso tiene una ventaja importante sobre la regresión Ridge, ya que produce modelos más simples y más interpretables que involucran solo un subconjunto de los predictores. Sin embargo, ¿qué método conduce a una mejor precisión de predicción? La figura VII muestra la varianza, el sesgo cuadrado y el MSE de prueba del lasso aplicado a los mismos datos simulados que en la figura V. Claramente, Lasso conduce a un comportamiento cualitativamente similar al de la regresión Ridge, en el sentido de que a medida que λ aumenta, la varianza disminuye y el sesgo aumenta. En el panel derecho de la figura VII, las líneas punteadas representan los ajustes de la regresión Ridge. Aquí se grafican ambos contra su R^2 en los datos de entrenamiento. Esta es otra forma útil de indexar modelos, y puede usarse para comparar modelos con diferentes tipos de regularización, como es el caso aquí. En este ejemplo, Lasso y la regresión Ridge resultan en sesgos casi idénticos. Sin embargo, la varianza de la regresión Ridge es ligeramente menor que la varianza del lasso. En consecuencia, el MSE mínimo de la regresión Ridge es ligeramente menor que el de Lasso.

Sin embargo, los datos en la figura VII fueron generados de tal manera que todos los 45 predictores estaban relacionados con la respuesta, es decir, ninguno de los coeficientes verdaderos $\beta_1, \ldots, \beta_{45}$ era igual a cero. Lasso asume implícitamente que varios de los coeficientes son realmente iguales a cero. En consecuencia, no es sorprendente que la regresión Ridge supere a Lasso en términos de error de predicción en este contexto. La Figura VIII ilustra una situación similar, excepto que ahora la respuesta es una función de solo 2 de los 45 predictores. Ahora, Lasso tiende a mejorar la actuación de la regresión Ridge en términos de sesgo, varianza y MSE.

Estos dos ejemplos ilustran que ni la regresión Ridge ni Lasso dominarán al otro. En general, uno podría esperar que Lasso funcione mejor en un contexto donde un número

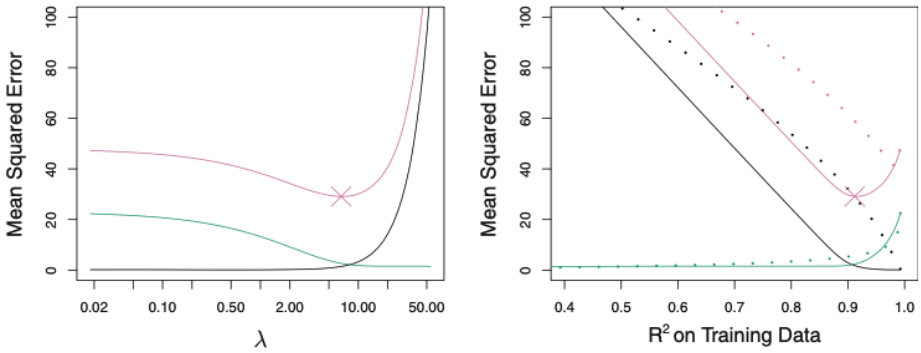

FIGURE VIII: Izquierda: Gráficas del sesgo cuadrado (negro), varianza (verde) y MSE de prueba (morado) para Lasso en un conjunto de datos simulado. Derecha: Comparación del sesgo cuadrado, varianza y MSE de prueba entre Lasso (sólido) y Ridge (discontinuo). Ambos se grafican contra su R^2 en los datos de entrenamiento, como una forma común de indexación. Las cruces en ambas gráficas indican el modelo Lasso para el cual el MSE es el más pequeño.

relativamente pequeño de predictores tenga coeficientes sustanciales, y los predictores restantes tengan coeficientes muy pequeños o iguales a cero. La regresión Ridge funcionará mejor cuando la respuesta sea una función de muchos predictores, todos con coeficientes de tamaño aproximadamente igual. Sin embargo, el número de predictores que está relacionado con la respuesta nunca se conoce a priori para conjuntos de datos reales. Una técnica como la validación cruzada puede usarse para determinar qué enfoque es mejor en un conjunto de datos particular.

Al igual que en la regresión Ridge, cuando las estimaciones de mínimos cuadrados tienen una varianza excesivamente alta, la solución de Lasso puede producir una reducción en la varianza a expensas de un pequeño aumento en el sesgo y, en consecuencia, puede generar predicciones más precisas. A diferencia de la regresión Ridge, Lasso realiza una selección de variables y, por lo tanto, resulta en modelos que son más fáciles de interpretar.

2.4.3 Selección del parámetro de ajuste

Al igual que los enfoques de selección de subconjuntos considerados en la sección 4.1 requieren un método para determinar cuál de los modelos en consideración es el mejor, implementar la regresión Ridge y Lasso requiere un método para seleccionar un valor para el parámetro de ajuste λ en (2.35) y (2.40), o, de manera equivalente, el valor de la restricción s en (2.37) y (2.42). La validación cruzada proporciona una forma sencilla de abordar este problema. Se elige una cuadrícula de valores de λ y se calcula el error de validación cruzada para cada valor de λ. Luego se selecciona el valor del parámetro de ajuste para el cual el error de validación cruzada es el más pequeño. Finalmente, el modelo se vuelve a ajustar utilizando todas las observaciones disponibles y el valor seleccionado del parámetro de ajuste.

FALTA PROBLEMAs en alta dimension !!!!

2.5 Reducción de dimensión

El análisis de componentes principales (PCA) es un enfoque popular para derivar un conjunto de características de baja dimensión a partir de un gran conjunto de variables. Aquí se describe su uso como una técnica de reducción de dimensión para la regresión.

2.5.1 Análisis de componentes principales

Una única muestra

Para un vector aleatorio $X = (X_1, \ldots, X_p)^T$, el vector de medias es $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^T$, y la matriz de covarianza de la población Σ viene dada por

$$\Sigma = \begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_p) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) & \cdots & \mathrm{Cov}(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_p, X_1) & \mathrm{Cov}(X_p, X_2) & \cdots & \mathrm{Var}(X_p) \end{pmatrix} \tag{2.43}$$

Sean $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ los autovalores de Σ, y $\mathbf{e}_1, \ldots, \mathbf{e}_p$ los correspondientes vectores propios (debidamente normalizados). La varianza de la componente principal k-ésima será igual a λ_k, mientras que los elemtnos de $\mathbf{e}_k$ serán los coeficientes de la componente principal k-ésima. La primera componente principal será la que mayor varianza tenga, y su autovalor será el mayor.

La proporción de varianza explicada por la componente principal k-ésima es

$$\frac{\lambda_k}{\sum_{j=1}^{p} \lambda_j} \tag{2.44}$$

Por tanto, la proporción de varianza explicada por las M primeras componentes principales es

$$\frac{\sum_{j=1}^{M} \lambda_j}{\sum_{j=1}^{p} \lambda_j} \tag{2.45}$$

La matriz de correlación será

$$\rho = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{pmatrix}, \quad \rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j} = \frac{\mathrm{Cov}(X_i, X_j)}{\sqrt{\mathrm{Var}(X_i)\mathrm{Var}(X_j)}} \tag{2.46}$$

donde ρ_{ij} es la correlación entre las variables X_i y X_j

n muestras

Sea un conjunto de n observaciones y p características $X_1, X_2, \ldots, X_p$,

$$X = \begin{pmatrix} x11 & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \tag{2.47}$$

Las componentes de este conjunto suelen ser dependientes y contienen información redundante. Por tanto, podría resultar útil tomar combinaciones lineales de las variables originales, manteniendo la mayor cantidad de información posible. Cada una de las n observaciones yace en un espacio de p dimensiones, y el objetivo del PCA es proyectar estos datos en un espacio de menor dimensión que contenga la mayor cantidad de variabilidad en los datos.

Si ahora se toma un conjunto de n observaciones y p características (2.47), se puede estimar $\boldsymbol{\mu}$ como $\bar{x} = (\bar{x}_1, \ldots, \bar{x}_p)^T$, y la matriz de covarianza como

$$S = \begin{pmatrix} s_1^2 & s_{12} & \cdots & s_{1d} \\ s_{21} & s_2^2 & \cdots & s_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ s_{d1} & s_{d2} & \cdots & s_d^2 \end{pmatrix} \tag{2.48}$$

El análisis de componentes principales se centra en explicar la estructura de varianza-covarianza de $X = (X_1, \ldots, X_p)^T$ a partir de un conjunto menor de variables no correlacionadas (las componentes principales). Las componentes principales Z son combinaciones lineales no correlacionadas de las características $X_1, \ldots, X_p$ y cuya varianza es lo más grande posible. Matemáticamente, las componentes principales son

$$Z_m = \phi_{m1} X_1 + \phi_{m2} X_2 + \cdots + \phi_{mp} X_p = \boldsymbol{\phi}_m^T X \tag{2.49}$$

maximizando $\text{Var}(Z_1) = \boldsymbol{\phi}_m^T S \boldsymbol{\phi}_m$ sujeto a $||\boldsymbol{\phi}_m|| = \sum_{j=1}^{p} \phi_{jm}^2 = 1$ y $\text{Cov}(Z_k, Z_m) = \boldsymbol{\phi}_k^T S \boldsymbol{\phi}_m = 0$ para $m < k$.

En este caso, se puede estimar la matriz de covarianza con la matriz de correlación muestral,

$$R = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix}, \quad r_{ij} = \frac{s_{ij}}{s_i s_j} \tag{2.50}$$

2.5.2 Regresión de componentes principales

El análisis de componentes principales puede usarse como método de reducción de dimensión en problemas de regresión. Sea una respuesta cuantitativa Y y un conjunto de predictores $\mathbf{X} = (X_1, \ldots, X_p)^T$. La regresión de componentes principales se basa en construir las primeras M componentes principales $Z_1, \ldots, Z_M$ y luego usarlas como predictores en un modelo de regresión lineal ajustado con mínimos cuadrados,

$$Y = \beta + \sum_{i=1}^{M} \beta_i Z_i + \epsilon \tag{2.51}$$

Esta técnica asume que las direcciones en las que $X_1, \ldots, X_p$ muestran mayor variabilidad son las direcciones asociadas con Y, las más importantes a la hora de predecirla. La estimación de $k \ll p$ puede resolver problemas de *overfitting*.

2.5.3 Consideraciones en PCA

Escalado de variables

Antes de realizar el PCA, las variables deben centrarse para tener una media de cero (la media por columnas de $\mathbf{X}$ debe ser 0). Además, los resultados obtenidos al realizar el PCA también dependerán de si las variables se han escalado individualmente (cada una multiplicada por una constante diferente). Esto contrasta con algunas otras técnicas de aprendizaje supervisado y no supervisado, como la regresión lineal, en las que escalar las variables no tiene efecto.

Además, resulta poco deseable que las componentes principales obtenidas dependan de una elección arbitraria de escala, por lo que se suele escalar cada variable para tener desviación estándar unitaria antes de realizar el PCA. Sin embargo, si las variables están medidas en las mismas unidades, este escalado no es necesario.

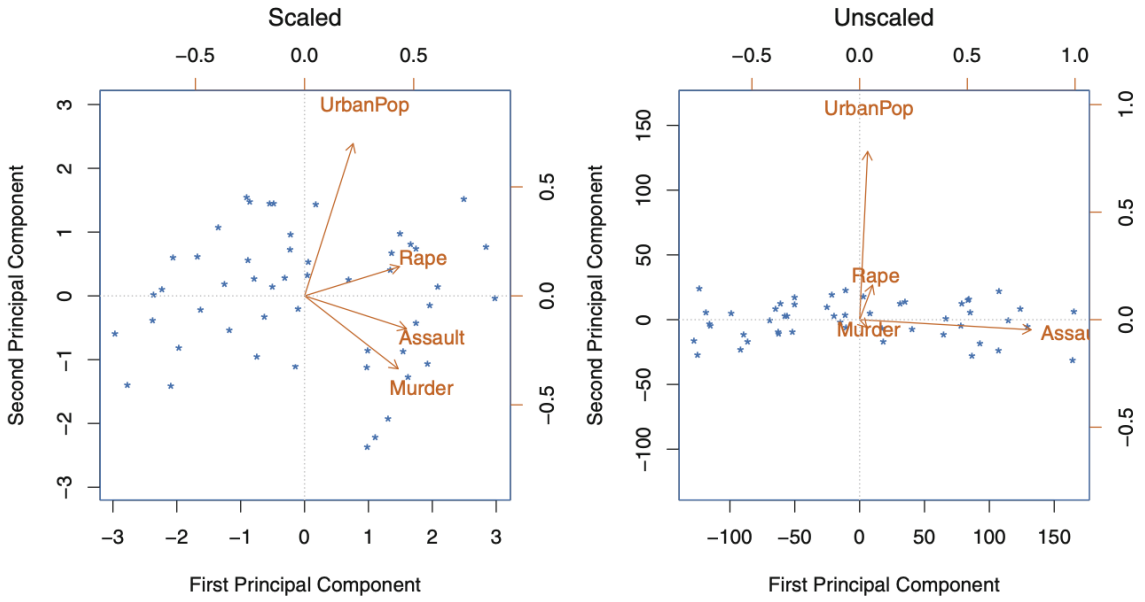

FIGURE IX: Dos biplots de componentes principales para los datos de *USArrests*. Variables escaladas para tener desviaciones estándar unitarias. Derecha: componentes principales usando datos no escalados. *Assault* tiene, con diferencia, la mayor carga en el primer componente principal porque tiene la mayor varianza entre las cuatro variables. En general, se recomienda escalar las variables para que tengan una desviación estándar de uno.

Unicidad de las componentes principales

Cada vector de cargas de los componentes principales es único, salvo un cambio de signo. Los signos pueden diferir porque cada vector de cargas de los componentes principales especifica una dirección en el espacio p-dimensional: cambiar el signo no tiene efecto ya que la dirección no cambia.

3 Clasificacion

El modelo de regresión lineal presentado en el capítulo 3 asume que la variable de respuesta Y es cuantitativa. Sin embargo, en muchas situaciones, la variable de respuesta es cualitativa. A menudo las variables cualitativas se denominan categóricas. Los enfoques para predecir respuestas cualitativas son proceso conocidos como clasificación, ya que implica asignar la observación a una categoría o clase. Por otro lado, frecuentemente los métodos utilizados para clasificación primero predicen la probabilidad de cada una de las categorías de una variable cualitativa, como base para realizar la clasificación. En este sentido, también se comportan como métodos de regresión.

3.1 El entorno de clasificación

Muchos de los conceptos encontrados en los capítulos anteriores, como el equilibrio entre sesgo y varianza, se transfieren al contexto de clasificación con solo algunas modificaciones debido a que y_i ya no es numérica. Supongamos que se busca estimar f basándose en observaciones de entrenamiento $(x_1, y_1), \ldots, (x_n, y_n)$, donde ahora $y_1, ..., y_n$ son cualitativas. El enfoque más común para cuantificar la precisión de nuestra estimación $\hat{f}$ es la tasa de error de entrenamiento, la proporción de errores que se cometen si se aplica la estimación $\hat{f}$ a las observaciones de entrenamiento:

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i) \tag{3.1}$$

Aquí $\hat{y}_i$ es la etiqueta de clase predicha para la i-ésima observación usando $\hat{f}$, y $I(y_i \neq \hat{y}_i)$ es una variable indicadora que equivale a 1 si $y_i \neq \hat{y}_i$ y cero si $y_i = \hat{y}_i$. Si $I(y_i \neq \hat{y}_i) = 0$ entonces la i-ésima observación fue clasificada correctamente por nuestro método de clasificación; de lo contrario fue mal clasificada. Por lo tanto, la ecuación 3.1 calcula la fracción de clasificaciones incorrectas.

La ecuación 3.1 se conoce como la tasa de error de entrenamiento porque se calcula basándose en los datos que se utilizaron para entrenar el clasificador. Como en el contexto de regresión, se está más interesado en las tasas de error que resultan de aplicar nuestro clasificador a observaciones de prueba que no se utilizaron en el entrenamiento. La tasa de error de prueba asociada con un conjunto de observaciones de prueba de la forma (x_0, y_0) está dada por:

$$\text{AVE}(I(y_0 \neq \hat{y}_0)) \tag{3.2}$$

donde $\hat{y}_0$ es la etiqueta de clase predicha que resulta de aplicar el clasificador a la observación de prueba con predictor x_0. Un buen clasificador es aquel para el cual el error de prueba (3.2) es mínimo.

3.1.1 El clasificador de Bayes

Se puede demostrar que la tasa de error de prueba dada en (3.2) se minimiza, en promedio, por un clasificador muy simple que asigna cada observación a la clase más probable, dados sus valores de predictores. En otras palabras, se debe simplemente asignar una observación de prueba con vector de predictores x_0 a la clase j para la cual

$$\Pr(Y = j|X = x_0) \tag{3.3}$$

es mayor. Nótese que (3.3) es una probabilidad condicional: es la probabilidad de que $Y = j$, dado el vector de predictores observado x_0. Este clasificador tan simple se llama clasificador de Bayes. En un problema de dos clases donde solo hay dos posibles valores de respuesta, digamos clase 1 o clase 2, el clasificador de Bayes corresponde a predecir la clase uno si $\Pr(Y = 1|X = x_0) > 0.5$, y la clase dos en caso contrario.

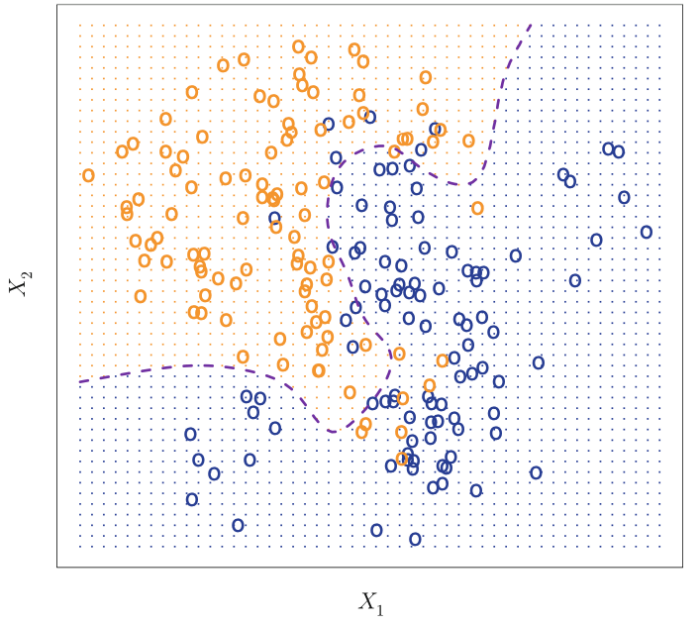

FIGURE X: Conjunto de datos simulado de 100 observaciones en cada uno de los dos grupos, indicados en azul y en naranja. La línea discontinua púrpura representa la frontera de decisión de Bayes. La cuadrícula de fondo naranja indica la región en la cual una observación de prueba será asignada a la clase naranja, y la cuadrícula de fondo azul indica la región en la cual una observación de prueba será asignada a la clase azul.

La figura X proporciona un ejemplo usando un conjunto de datos simulado en un espacio bidimensional que consiste en los predictores X_1 y X_2. Los círculos naranjas y azules corresponden a observaciones de entrenamiento que pertenecen a dos clases diferentes. Para cada valor de X_1 y X_2, hay una probabilidad diferente de que la respuesta sea naranja o azul. Dado que estos son datos simulados, se sabe cómo se generaron los datos y se pueden calcular las probabilidades condicionales para cada valor de X_1 y X_2. La región sombreada en naranja refleja el conjunto de puntos para los cuales $\Pr(Y = \text{naranja}|X)$ es mayor al 50%, mientras que la región sombreada en azul indica el conjunto de puntos para los cuales la probabilidad es menor al 50%. La línea discontinua púrpura representa los puntos donde la probabilidad es exactamente del 50%. Esto se llama la frontera de decisión de Bayes. La predicción del

clasificador de Bayes está determinada por la frontera de decisión de Bayes; una observación que cae en el lado naranja de la frontera se asignará a la clase naranja, y de manera similar, una observación en el lado azul de la frontera se asignará a la clase azul.

El clasificador de Bayes produce la tasa de error de prueba más baja posible, llamada la tasa de error de Bayes. Dado que el clasificador de Bayes siempre elegirá la clase para la cual (3.3) es mayor, la tasa de error en $X = x_0$ será

$$1 - \max_j \Pr(Y = j|X = x_0) \tag{3.4}$$

En general, la tasa de error de Bayes global está dada por

$$1 - E\left(\max_j \Pr(Y = j|X)\right) \tag{3.5}$$

donde la expectativa promedia la probabilidad sobre todos los valores posibles de X. Para nuestros datos simulados, la tasa de error de Bayes es 0.1304. Es mayor que cero, porque las clases se superponen en la población verdadera, por lo que $\max_j \Pr(Y = j|X = x_0) < 1$ para algunos valores de x_0. La tasa de error de Bayes es análoga al error irreducible discutido anteriormente.

3.1.2 ¿Por qué no regresión lineal?

Supongamos que se está tratando de predecir la condición médica de un paciente en la sala de emergencias en función de sus síntomas. En este ejemplo simplificado, hay tres posibles diagnósticos: derrame cerebral, sobredosis de drogas y ataque epiléptico. Se podría considerar codificar estos valores como una variable de respuesta cuantitativa, Y, de la siguiente manera:

$$Y = \begin{cases} 0 & \text{si derrame cerebral} \\ 1 & \text{si sobredosis de drogas} \\ 2 & \text{si ataque epiléptico} \end{cases}$$

Usando esta codificación, se podría usar mínimos cuadrados para ajustar un modelo de regresión lineal para predecir Y en función de un conjunto de predictores $X_1, ..., X_p$. Desafortunadamente, esta codificación implica un orden en los resultados, colocando sobredosis de drogas entre derrame cerebral y ataque epiléptico, e insistiendo en que la diferencia entre derrame cerebral y sobredosis de drogas es la misma que la diferencia entre sobredosis de drogas y ataque epiléptico. En la práctica, no hay ninguna razón particular para que esto sea así. Por ejemplo, se podría elegir una codificación igualmente razonable,

$$Y = \begin{cases} 0 & \text{si sobredosis de drogas} \\ 1 & \text{si derrame cerebral} \\ 2 & \text{si ataque epiléptico} \end{cases}$$

lo que implicaría una relación totalmente diferente entre las tres condiciones. Cada una de estas codificaciones produciría modelos lineales fundamentalmente diferentes que llevarían a diferentes conjuntos de predicciones en observaciones de prueba.

Si los valores de la variable de respuesta tuvieran un orden natural, como leve, moderado y severo, y se sintiera que la brecha entre leve y moderado es similar a la brecha entre moderado

y severo, entonces una codificación 1, 2, 3 sería razonable. Desafortunadamente, en general no hay una manera natural de convertir una variable de respuesta cualitativa con más de dos niveles en una respuesta cuantitativa que esté lista para la regresión lineal.

Para una respuesta cualitativa binaria (de dos niveles), la situación es mejor. Por ejemplo, tal vez solo hay dos posibilidades para la condición médica del paciente: derrame cerebral y sobredosis de drogas. Entonces se podría potencialmente usar el enfoque de variable ficticia para codificar la respuesta de la siguiente manera:

$$Y = \begin{cases} 0 & \text{si derrame cerebral} \\ 1 & \text{si sobredosis de drogas} \end{cases}$$

Se podría ajustar una regresión lineal a esta respuesta binaria, y predecir sobredosis de drogas si $\hat{Y} > 0.5$ y derrame cerebral en caso contrario. En el caso binario, no es difícil mostrar que incluso si se invierte la codificación anterior, la regresión lineal producirá las mismas predicciones finales.

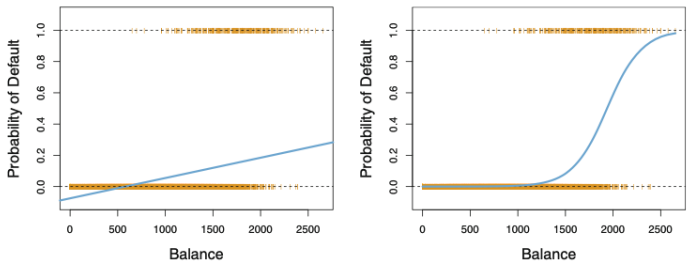

FIGURE XI: Clasificación usando los datos de Default. Izquierda: Probabilidad estimada de incumplimiento usando regresión lineal. ¡Algunas probabilidades estimadas son negativas! Las marcas naranjas indican los valores 0/1 codificados para incumplimiento (No o Sí). Derecha: Probabilidades predichas de incumplimiento usando regresión logística. Todas las probabilidades están entre 0 y 1.

Para una respuesta binaria con una codificación 0/1 como la anterior, la regresión por mínimos cuadrados tiene sentido; se puede mostrar que el $\hat{X}\beta$ obtenido usando regresión lineal es de hecho una estimación de Pr(sobredosis de drogas$|X$) en este caso especial. Sin embargo, si se usa regresión lineal, algunas de nuestras estimaciones podrían estar fuera del intervalo [0,1] (ver figura XI), lo que las hace difíciles de interpretar como probabilidades. No obstante, las predicciones proporcionan un orden y pueden interpretarse como estimaciones de probabilidad crudas. Curiosamente, resulta que las clasificaciones que se obtienen si se usa regresión lineal para predecir una respuesta binaria serán las mismas que para el procedimiento de análisis discriminante lineal (LDA).

Sin embargo, el enfoque de variable ficticia no se puede extender fácilmente para acomodar respuestas cualitativas con más de dos niveles. Por estas razones, es preferible usar un método de clasificación que esté verdaderamente adaptado para valores de respuesta cualitativos, como los que se presentan a continuación.

3.2 Regresión logística

3.2.1 Modelo logístico

Veamos cómo se debe modelar la relación entre $p(X) = \Pr(Y = 1|X)$ y X (por conveniencia se usará la codificación genérica 0/1 para la respuesta). Anteriormentese se habló de usar un modelo de regresión lineal para representar estas probabilidades:

$$p(X) = \beta_0 + \beta_1 X \tag{3.6}$$

Si se usa este enfoque para predecir default=Sí usando balance, entonces se obtiene el modelo mostrado en el panel izquierdo de la figura XI. Cada vez que se ajusta una línea recta a una respuesta binaria que está codificada como 0 o 1, en principio siempre se puede predecir $p(X) < 0$ para algunos valores de X y $p(X) > 1$ para otros (a menos que el rango de X esté limitado).

Para evitar este problema, se debe modelar $p(X)$ usando una función que dé salidas entre 0 y 1 para todos los valores de X. Nótese que la frontera de decisión entre ambas salidas viene dada por $P(Y = 1|X) = P(Y = 0|X)$. Muchas funciones cumplen con esta descripción. En la regresión logística, se usa la función logística,

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}} = \frac{1}{1 + e^{-(\beta_0+\beta_1 X)}} \tag{3.7}$$

Para ajustar el modelo (3.7), se usa un método llamado máxima verosimilitud, que se discute en la siguiente sección. El panel derecho de la figura XI ilustra el ajuste del modelo de regresión logística a los datos de Default. Nótese que para balances bajos ahora se predice la probabilidad de incumplimiento como cercana a cero, pero nunca por debajo. Del mismo modo, para balances altos se predice una probabilidad de incumplimiento cercana a uno, pero nunca por encima. Después de manipular un poco (3.7), se encuentra que

$$\frac{p(X)}{1 - p(X)} = e^{\beta_0+\beta_1 X} \tag{3.8}$$

La cantidad $\frac{p(X)}{1-p(X)}$ se llama las probabilidades (*odds*), y puede tomar cualquier valor entre 0 e ∞. Valores de las probabilidades cercanos a 0 e ∞ indican probabilidades muy bajas y muy altas de incumplimiento, respectivamente. Al tomar el logaritmo de ambos lados de (3.8), se llega a

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X \tag{3.9}$$

La transformación monótona del lado izquierdo se llama el logaritmo de las probabilidades (*log-odds*) o *logit*. Se ve que el modelo de regresión logística (3.7) tiene un *logit* que es lineal en X.

En un modelo de regresión lineal, β_1 da el cambio promedio en Y asociado con un aumento de una unidad en X. En contraste, en un modelo de regresión logística, aumentar X en una unidad cambia el logaritmo de las probabilidades en β_1 (3.9) o, equivalentemente, multiplica las probabilidades por e^{β_1} (3.8). Sin embargo, debido a que la relación entre $p(X)$ y X en (3.7)) no es una línea recta, β_1 no corresponde al cambio en $p(X)$ asociado con un aumento de una unidad en X. La cantidad que $p(X)$ cambia debido a un cambio de una unidad en X dependerá del valor actual de X. Pero independientemente del valor de X, si β_1 es positivo,

entonces aumentar X estará asociado con un aumento en $p(X)$, y si β_1 es negativo, entonces aumentar X estará asociado con una disminución en $p(X)$.

3.2.2 Estimación de los coeficientes

Los coeficientes β_0 y β_1 en (3.7) son desconocidos y deben ser estimados basándose en los datos de entrenamiento disponibles. Aunque se podría usar mínimos cuadrados (no lineales) para ajustar el modelo (3.9), el método de máxima verosimilitud es preferible, ya que tiene mejores propiedades estadísticas. La intuición básica detrás del uso de máxima verosimilitud para ajustar un modelo de regresión logística es la siguiente: se intenta encontrar $\hat{\beta}_0$ y $\hat{\beta}_1$ de manera que al insertar estas estimaciones en el modelo para $p(X)$, dado en (3.7), se obtenga un número cercano a uno para todos los individuos que cumplieron, y un número cercano a cero para todos los individuos que incumplieron. Esta intuición se puede formalizar usando una ecuación matemática llamada función de verosimilitud:

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1}^{n} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'})) \tag{3.10}$$

Las estimaciones $\hat{\beta}_0$ y $\hat{\beta}_1$ se eligen para maximizar esta función de verosimilitud. En el contexto de la regresión lineal, el enfoque de mínimos cuadrados es de hecho un caso especial de máxima verosimilitud.

Muchos aspectos de la salida de la regresión logística son similares a la salida de la regresión lineal. Por ejemplo, se puede medir la precisión de las estimaciones de los coeficientes calculando sus errores estándar. El estadístico z juega el mismo papel que el estadístico t en la salida de la regresión lineal. Por ejemplo, el estadístico z asociado con $\hat{\beta}_1$ es igual a $\hat{\beta}_1/SE(\hat{\beta}_1)$, por lo que un valor grande (absoluto) del estadístico z indica evidencia en contra de la hipótesis nula $H_0 : \beta_1 = 0$. Esta hipótesis nula implica que $p(X) = \frac{e^{\beta_0}}{1+e^{\beta_0}}$, en otras palabras, que la probabilidad de incumplimiento no depende del balance. Si el valor p asociado con balance es muy pequeño, se puede rechazar H_0. En otras palabras, se concluye que efectivamente hay una asociación entre balance y probabilidad de incumplimiento.

Una vez estimados los coeficientes, se pueden hacer predicciones de la probabilidad $\hat{p}(X)$ de forma sencilla

$$\hat{p}(X) = \frac{1}{1 + e^{-(\hat{\beta}_0 + \hat{\beta}_1 X)}} \tag{3.11}$$

3.2.3 Regresión logística múltiple

Sea el problea de predecir una respuesta binaria usando múltiples predictores. La regresión logística anterior se puede generalizar de forma inmediata, de modo que el logarímto de las probabilidades serán

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \tag{3.12}$$

y la probabilidad será

$$p(X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)}} \tag{3.13}$$

3.2.4 Regresión logística no binaria

Los modelos de regresión logística de binarios discutidos en las secciones anteriores tienen generalizaciones para múltiples clases, pero en la práctica no se utilizan con tanta frecuencia. Una de las razones es que el método que se discute en la próxima sección, el análisis discriminante, es popular para la clasificación de múltiples clases.

3.3 Análisis discriminante lineal

La regresión logística implica modelar directamente $\Pr(Y = k|X = x)$ usando la función logística, dada por (3.13) para el caso de dos clases de respuesta. Se modela la distribución condicional de la respuesta Y, dado el(los) predictor(es) X. Ahora se considera un enfoque alternativo y menos directo para estimar estas probabilidades. En este enfoque alternativo (modelo generativo), se modela la distribución de los predictores X por separado en cada una de las clases de respuesta (es decir, dado Y), y luego se usa el teorema de Bayes para convertir estas distribuciones en estimaciones de $\Pr(Y = k|X = x)$. Cuando se asume que estas distribuciones son normales, resulta que el modelo es muy similar en forma a la regresión logística. Hay varias razones para usar un método distinto a la regresión logística (modelo discriminante):

- Cuando las clases están bien separadas, las estimaciones de los parámetros para el modelo de regresión logística son sorprendentemente inestables. El análisis discriminante lineal no sufre de este problema.
- Si n es pequeño y la distribución de los predictores X es aproximadamente normal en cada una de las clases, el modelo de análisis discriminante lineal es nuevamente más estable que el modelo de regresión logística.
- El análisis discriminante lineal es popular cuando se tienen más de dos clases de respuesta.

3.3.1 Teorema de Bayes para clasificación

Regla de Bayes

Sean dos eventos cualesquiera A y B. La regla de Bayes establece que para encontrar $P(B|A)$ (probabilidad de que ocurra B dado que A ocurrió), se puede usar la siguiente relación:

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A|B)P(B)}{P(A)} \tag{3.14}$$

Regla de Bayes en problemas de clasificación

Supongamos que se desea clasificar una observación en una de K clases distintas, donde $K \geq 2$, es decir, la variable de respuesta cualitativa Y puede tomar K valores distintos y no ordenados. Sea π_k la probabilidad general o *a priori* de que una observación elegida al azar provenga de la k-ésima clase; esta es la probabilidad de que una observación dada esté asociada con la k-ésima categoría de la variable de respuesta Y. Sea $f_k(X) \equiv \Pr(X = x|Y = k)$ la función de densidad de X para una observación que proviene de la k-ésima clase. Así, $f_k(x)$ es relativamente grande si hay una alta probabilidad de que una observación en la k-ésima clase tenga $X \approx x$, y es pequeña si es muy improbable que una observación en la k-ésima clase tenga $X \approx x$.

Entonces, el teorema de Bayes establece que

$$\Pr(Y = k|X = x) = \frac{P(X = x|\ Y = k)P(Y = k)}{P(X = x)} = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)} \tag{3.15}$$

Se usará la abreviatura $p_k(X) = \Pr(Y = k|X = x)$. Esto sugiere que en lugar de calcular directamente $p_k(X)$, simplemente se pueden insertar estimaciones de π_k y $f_k(X)$ en (3.15). En general, estimar π_k es fácil si se tiene una muestra aleatoria de Ys de la población: simplemente se calcula la fracción de las observaciones de entrenamiento que pertenecen a la k-ésima clase. Sin embargo, estimar $f_k(X)$ tiende a ser más complicado, a menos que se asuman algunas formas simples para estas densidades. Se refiere a $p_k(x)$ como la probabilidad posterior de que una observación $X = x$ pertenezca a la k-ésima clase. Es decir, es la probabilidad de que la observación pertenezca a la k-ésima clase, dado el valor del predictor para esa observación.

Se sabe que el clasificador de Bayes, que clasifica una observación a la clase para la cual $p_k(X)$ es mayor, tiene la tasa de error más baja posible entre todos los clasificadores. (Esto, por supuesto, solo es cierto si los términos en (3.15) están todos especificados correctamente). Por lo tanto, si se puede encontrar una manera de estimar $f_k(X)$, entonces se puede desarrollar un clasificador que aproxime al clasificador de Bayes. Esto se verá en las próximas secciones.

Análisis discriminante lineal para $p = 1$

Por ahora, supongamos que $p = 1$, es decir, solo se tiene un predictor. Se desea obtener una estimación para $f_k(x)$ que se pueda insertar en (3.15) para estimar $p_k(x)$. Luego se clasificará una observación en la clase para la cual $p_k(x)$ sea mayor. Para estimar $f_k(x)$, primero se harán algunas suposiciones sobre su forma.

Supongamos que $f_k(x)$ es normal o gaussiana. En el entorno unidimensional, la densidad normal toma la forma

$$f_k(x) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(x-\mu_k)^2}{2\sigma_k^2}\right) \tag{3.16}$$

donde μ_k y σ_k^2 son los parámetros de media y varianza para la k-ésima clase. Por ahora, supongamos además que $\sigma_1^2 = \ldots = \sigma_K^2$, es decir, hay un término de varianza compartido entre todas las K clases, que por simplicidad se puede denotar como σ^2. Insertando (3.16) en (3.15), se encuentra que

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu_k)^2}{2\sigma^2}\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu_l)^2}{2\sigma^2}\right)} \tag{3.17}$$

Nótese que en (3.17), π_k denota la probabilidad a priori de que una observación pertenezca a la k-ésima clase. El clasificador de Bayes implica asignar una observación $X = x$ a la clase para la cual (3.17) es mayor. Tomando el logaritmo de (3.17) y reorganizando los términos, no es difícil mostrar que esto es equivalente a asignar la observación a la clase para la cual

$$\delta_k(x) = x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k) \tag{3.18}$$

es mayor. Por ejemplo, si $K = 2$ y $\pi_1 = \pi_2$, entonces el clasificador de Bayes asigna una observación a la clase 1 si $2x(\mu_1 - \mu_2) > \mu_2^2 - \mu_1^2$, y a la clase 2 en caso contrario. En este

caso, la frontera de decisión de Bayes corresponde al punto donde

$$x = \frac{\mu_1^2 - \mu_2^2}{2(\mu_1 - \mu_2)} = \frac{\mu_1 + \mu_2}{2} \tag{3.19}$$

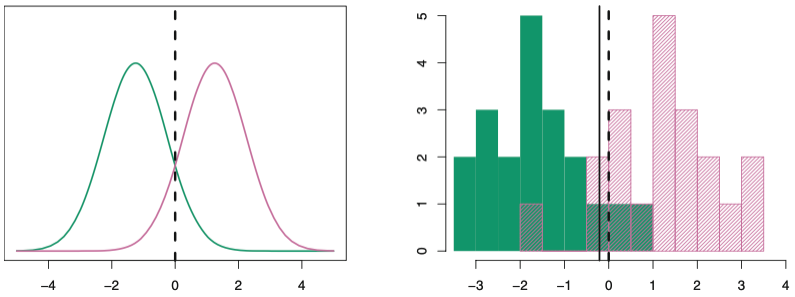

FIGURE XII: Izquierda: Se muestran dos funciones de densidad normal unidimensionales. La línea vertical discontinua representa la frontera de decisión de Bayes. Derecha: Se extrajeron 20 observaciones de cada una de las dos clases, y se muestran como histogramas. La frontera de decisión de Bayes se muestra nuevamente como una línea vertical discontinua. La línea vertical sólida representa la frontera de decisión LDA estimada a partir de los datos de entrenamiento.

Un ejemplo se muestra en el panel izquierdo de la figura XII. Las dos funciones de densidad normal que se muestran, $f_1(x)$ y $f_2(x)$, representan dos clases distintas. Los parámetros de media y varianza para las dos funciones de densidad son $\mu_1 = -1.25$, $\mu_2 = 1.25$, y $\sigma_1^2 = \sigma_2^2 = 1$. Las dos densidades se superponen, por lo que dado que $X = x$, hay cierta incertidumbre sobre la clase a la que pertenece la observación. Si se asume que una observación es igualmente probable que provenga de cualquiera de las dos clases, es decir, $\pi_1 = \pi_2 = 0.5$, entonces al inspeccionar (3.19), se ve que el clasificador de Bayes asigna la observación a la clase 1 si $x < 0$ y a la clase 2 en caso contrario.

Nótese que en este caso, se puede calcular el clasificador de Bayes porque se sabe que X se extrae de una distribución gaussiana dentro de cada clase, y se conocen todos los parámetros involucrados. En una situación de la vida real, no se puede calcular el clasificador de Bayes. En la práctica, incluso si se está bastante seguro de la suposición de que X se extrae de una distribución gaussiana dentro de cada clase, aún se deben estimar los parámetros $\mu_1, \ldots, \mu_K$, $\pi_1, \ldots, \pi_K$, y σ^2. El método de análisis discriminante lineal (LDA) aproxima el clasificador de Bayes insertando estimaciones para π_k, μ_k, y σ^2 en (3.18). En particular, se usan las siguientes estimaciones:

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i, \quad \hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2 \tag{3.20}$$

donde n es el número total de observaciones de entrenamiento, y n_k es el número de observaciones de entrenamiento en la k-ésima clase. La estimación para μ_k es simplemente el promedio de todas las observaciones de entrenamiento de la k-ésima clase, mientras que $\hat{\sigma}^2$ se puede ver como un promedio ponderado de las varianzas muestrales para cada una de las K clases. A veces se tiene conocimiento de las probabilidades de pertenencia a la clase $\pi_1, \ldots, \pi_K$, que se pueden usar directamente. En ausencia de cualquier información adicional, LDA estima π_k usando la proporción de las observaciones de entrenamiento que pertenecen a

la k-ésima clase. En otras palabras,

$$\hat{\pi}_k = \frac{n_k}{n} \tag{3.21}$$

El clasificador LDA inserta las estimaciones dadas en (3.20) y (3.21) en (3.18), y asigna una observación $X = x$ a la clase para la cual

$$\hat{delta}_k(x) = x\frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log(\hat{\pi}_k) \tag{3.22}$$

es mayor. La palabra "lineal" en el nombre del clasificador proviene del hecho de que las funciones discriminantes $\hat{\delta}_k(x)$ en (3.22) son funciones lineales de x (en lugar de una función más compleja de x).

El panel derecho de la figura XII muestra un histograma de una muestra aleatoria de 20 observaciones de cada clase. Para implementar LDA, se comenzó estimando π_k, μ_k, y σ^2 usando (3.20) y (3.21). Luego se calculó la frontera de decisión, mostrada como una línea sólida negra, que resulta de asignar una observación a la clase para la cual (3.22) es mayor. Todos los puntos a la izquierda de esta línea se asignarán a la clase verde, mientras que los puntos a la derecha de esta línea se asignarán a la clase púrpura. En este caso, dado que $n_1 = n_2 = 20$, se tiene $\hat{\pi}_1 = \hat{\pi}_2$. Como resultado, la frontera de decisión corresponde al punto medio entre las medias muestrales para las dos clases, $(\hat{\mu}_1 + \hat{\mu}_2)/2$. La figura indica que la frontera de decisión LDA está ligeramente a la izquierda de la frontera de decisión óptima de Bayes, que en cambio es igual a $(\mu_1 + \mu_2)/2 = 0$. Dado que estos son datos simulados, se puede generar un gran número de observaciones de prueba para calcular la tasa de error de Bayes y la tasa de error de prueba de LDA. Estas son 10.6% y 11.1%, respectivamente. En otras palabras, la tasa de error del clasificador LDA es solo 0.5% por encima de la tasa de error más baja posible. Esto indica que LDA está funcionando bastante bien en este conjunto de datos.

En resumen, el clasificador LDA resulta de suponer que las observaciones dentro de cada clase provienen de una distribución normal con un vector de media específico de la clase y una varianza común σ^2, e insertar estimaciones para estos parámetros en el clasificador de Bayes. Más adelante, se considerará un conjunto de suposiciones menos estrictas, permitiendo que las observaciones en la k-ésima clase tengan una varianza específica de la clase, σ_k^2.

Análisis discriminante lineal para $p > 1$

Para extender el clasificador LDA al caso de múltiples predictores, se asume que $X = (X_1, X_2, \ldots, X_p)$ se extrae de una distribución gaussiana multivariante (o normal multivariante), con un vector de media específico de la clase y una matriz de covarianza común.

La distribución gaussiana multivariante asume que cada predictor individual sigue una distribución normal unidimensional, como en (3.16), con alguna correlación entre cada par de predictores. Dos ejemplos de distribuciones gaussianas multivariantes con $p = 2$ se muestran en la figura XIII. La altura de la superficie en cualquier punto particular representa la probabilidad de que tanto X_1 como X_2 caigan en una pequeña región alrededor de ese punto. La forma de campana se distorsionará si los predictores están correlacionados o tienen varianzas desiguales, como se ilustra en el panel derecho de la figura XIII. En esta situación, la base de la campana tendrá una forma elíptica, en lugar de circular. Para indicar que una variable aleatoria p-dimensional X tiene una distribución gaussiana multivariante, escribimos $X \sim N(\mu, \Sigma)$. Aquí $E(X) = \mu$ es la media de X (un vector con p componentes), y $\text{Cov}(X) = \Sigma$ es la matriz

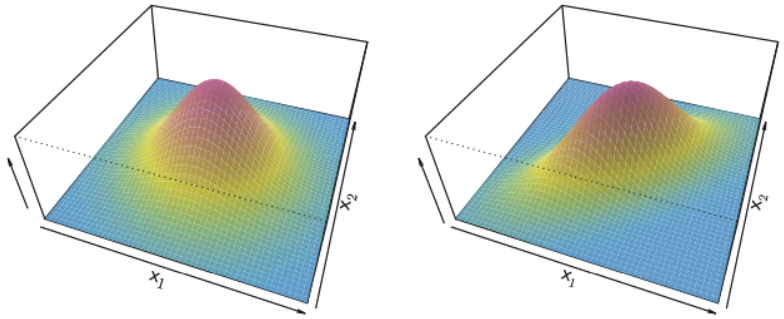

FIGURE XIII

de covarianza $p \times p$ de X. Formalmente, la densidad gaussiana multivariante se define como

$$f(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) \tag{3.23}$$

En el caso de más de un predictor ($p > 1$), el clasificador LDA asume que las observaciones en la k-ésima clase se extraen de una distribución gaussiana multivariante $N(\mu_k, \Sigma)$, donde μ_k es un vector de media específico de la clase, y Σ es una matriz de covarianza común a todas las K clases. Insertando la función de densidad para la k-ésima clase, $f_k(X = x)$, en (3.15) y realizando un poco de álgebra, se revela que el clasificador de Bayes asigna una observación $X = x$ a la clase para la cual

$$\delta_k(x) = x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \log(\pi_k) \tag{3.24}$$

es mayor. Esta es la versión vector/matriz de (3.18).

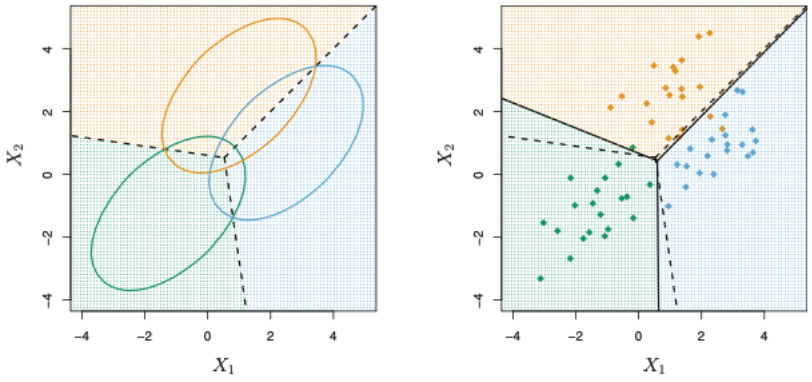

FIGURE XIV

Un ejemplo se muestra en el panel izquierdo de la figura XIV. Se muestran tres clases gaussianas de igual tamaño con vectores de media específicos de la clase y una matriz de covarianza común. Las tres elipses representan regiones que contienen el 95% de la probabilidad para cada una de las tres clases. Las líneas discontinuas son las fronteras de decisión de Bayes. En otras palabras, representan el conjunto de valores x para los cuales $\delta_k(x) = \delta_\ell(x)$; es decir,

$$x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k = x^T\Sigma^{-1}\mu_\ell - \frac{1}{2}\mu_\ell^T\Sigma^{-1}\mu_\ell \tag{3.25}$$

que puede reescribirse como

$$x^T\Sigma^{-1}(\mu_k - \mu_\ell) = \frac{1}{2}(\mu_k^T\Sigma^{-1}\mu_k - \mu_\ell^T\Sigma^{-1}\mu_\ell) \tag{3.26}$$

para $k \neq l$. (El término $\log(\pi_k)$ de (3.24) ha desaparecido porque cada una de las tres clases tiene el mismo número de observaciones de entrenamiento; es decir, π_k es el mismo para cada clase). Nótese que hay tres líneas que representan las fronteras de decisión de Bayes porque hay tres pares de clases entre las tres clases. Es decir, una frontera de decisión de Bayes separa la clase 1 de la clase 2, una separa la clase 1 de la clase 3, y una separa la clase 2 de la clase 3. Estas tres fronteras de decisión de Bayes dividen el espacio de predictores en tres regiones. El clasificador de Bayes clasificará una observación según la región en la que se encuentre.

Una vez más, se necesita estimar los parámetros desconocidos $\mu_1, \ldots, \mu_K$, $\pi_1, \ldots, \pi_K$, y Σ; las fórmulas son similares a las utilizadas en el caso unidimensional, dadas en (3.20). Para asignar una nueva observación $X = x$, LDA inserta estas estimaciones en (3.24) y clasifica a la clase para la cual $\hat{\delta}_k(x)$ es mayor. Nótese que en (3.24) $\delta_k(x)$ es una función lineal de x; es decir, la regla de decisión LDA depende de x solo a través de una combinación lineal de sus elementos. Una vez más, esta es la razón del término "lineal" en LDA.

En el panel derecho de la figura XIV, se muestran 20 observaciones extraídas de cada una de las tres clases, y las fronteras de decisión LDA resultantes se muestran como líneas negras sólidas. En general, las fronteras de decisión LDA están bastante cerca de las fronteras de decisión de Bayes, mostradas nuevamente como líneas discontinuas. Las tasas de error de prueba para los clasificadores de Bayes y LDA son 0.0746 y 0.0770, respectivamente. Esto indica que LDA está funcionando bien en estos datos. Las tasas de error de entrenamiento suelen ser más bajas que las tasas de error de prueba, que son la cantidad real de interés. La razón es que se ajustan específicamente los parámetros de nuestro modelo para que funcionen bien en los datos de entrenamiento. Cuanto mayor sea la relación de parámetros p con el número de muestras n, más se espera que este sobreajuste juegue un papel.

En la práctica, un clasificador binario como este puede cometer dos tipos de errores: puede asignar incorrectamente a un individuo que incumple a la categoría de no incumplimiento, o puede asignar incorrectamente a un individuo que no incumple a la categoría de incumplimiento. A menudo es de interés determinar cuál de estos dos tipos de errores se están cometiendo. Una matriz de confusión es una forma conveniente de mostrar esta información.

LDA está tratando de aproximar el clasificador de Bayes, que tiene la tasa de error total más baja de todos los clasificadores (si el modelo gaussiano es correcto). Es decir, el clasificador de Bayes producirá el menor número posible de observaciones mal clasificadas, independientemente de la clase de la que provengan los errores. Algunos errores de clasificación resultarán de asignar incorrectamente a un cliente que no incumple a la clase de incumplimiento, y otros resultarán de asignar incorrectamente a un cliente que incumple a la clase de no incumplimiento. En contraste, una compañía de tarjetas de crédito podría desear particularmente evitar clasificar incorrectamente a un individuo que incumplirá, mientras que clasificar incorrectamente a un individuo que no incumplirá, aunque aún debe evitarse, es menos problemático. Es posible modificar LDA para desarrollar un clasificador que satisfaga mejor las necesidades de la compañía de tarjetas de crédito.

El clasificador de Bayes funciona asignando una observación a la clase para la cual la probabilidad posterior $p_k(X)$ es mayor. En el caso de dos clases, esto equivale a asignar una observación a la clase *default* si

$$\Pr(\text{default} = \text{Yes}|X) > 0.5 \tag{3.27}$$

Por lo tanto, el clasificador de Bayes, y por extensión LDA, usa un umbral del 50% para la probabilidad posterior de *default* para asignar una observación a la clase de incumplimiento. Sin embargo, si preocupa predecir incorrectamente el estado de incumplimiento para los individuos que incumplen, entonces se puede considerar bajar este umbral. Por ejemplo, se podría etiquetar a cualquier cliente con una probabilidad posterior de incumplimiento superior al 20% a la clase de incumplimiento. En otras palabras, en lugar de asignar una observación a la clase *default* si (3.27) se cumple, se podría asignar una observación a esta clase si

$$\Pr(\text{default} = \text{Yes}|X) > 0.2 \tag{3.28}$$

Usar un umbral de 0.5, como en (3.27), minimiza la tasa de error general. Esto es de esperar, ya que el clasificador de Bayes usa un umbral de 0.5 y se sabe que tiene la tasa de error general más baja. Pero cuando se usa un umbral de 0.5, la tasa de error entre los individuos que incumplen es bastante alta (línea discontinua azul). A medida que se reduce el umbral, la tasa de error entre los individuos que incumplen disminuye constantemente, pero la tasa de error entre los individuos que no incumplen aumenta.

La curva ROC es un gráfico popular para mostrar simultáneamente los dos tipos de errores para todos los umbrales posibles. El nombre "ROC" es histórico y proviene de la teoría de comunicaciones. Es un acrónimo de características operativas del receptor. El rendimiento general de un clasificador, resumido en todos los umbrales posibles, se da por el área bajo la curva (AUC). Una curva ROC ideal abrazará la esquina superior izquierda, por lo que cuanto mayor sea el AUC, mejor será el clasificador. Se espera que un clasificador que no funcione mejor que el azar tenga un AUC de 0.5 (cuando se evalúa en un conjunto de prueba independiente no utilizado en el entrenamiento del modelo). Las curvas ROC son útiles para comparar diferentes clasificadores, ya que tienen en cuenta todos los umbrales posibles.

True class	Clase predicha		
	− o nulo	+ or no nulo	Total
− o nulo	True Neg. (TN)	False Pos. (FP)	N
+ o no nulo	False Neg. (FN)	True Pos. (TP)	P
Total	N*	P*	

TABLE 3.1: Matriz de confusión. Posibles resultados al aplicar un clasificador a una población.

Variar el umbral del clasificador cambia su tasa de verdaderos positivos y su tasa de falsos positivos. Estas también se llaman la sensibilidad y uno menos la especificidad del clasificador. Dado que hay una variedad casi desconcertante de términos utilizados en este contexto, ahora damos un resumen. La tabla 3.1 muestra los posibles resultados al aplicar un clasificador (o prueba de diagnóstico) a una población.

La tabla 3.2 enumera muchas de las medidas de rendimiento populares que se utilizan en este contexto. Los denominadores para las tasas de falsos positivos y verdaderos positivos son

Nombre	Definición	Sinónimos
Tasa de verdaderos positivos	TP/P	1 - Error de tipo 2, sensibilidad, recall, probabilidad de detección
Tasa de falsos positivos	FP/N	Error de tipo 1, 1 - especificidad
Valor predictivo positivo	TP/P^*	Precisión, 1 - proporcion de falsos descumbrimientos
Valor predictivo negativo	TN/N^*	

TABLE 3.2: Medidas de rendimiento para clasificación.

los conteos de población reales en cada clase. En contraste, los denominadores para el valor predictivo positivo y el valor predictivo negativo son los conteos totales predichos para cada clase.

Análisis discriminante cuadrático

Como se ha discutido, LDA asume que las observaciones dentro de cada clase se extraen de una distribución gaussiana multivariante con un vector de media específico de la clase y una matriz de covarianza común a todas las K clases. El análisis discriminante cuadrático (QDA) proporciona un enfoque alternativo. Al igual que LDA, el clasificador QDA resulta de asumir que las observaciones de cada clase se extraen de una distribución gaussiana, e insertar estimaciones para los parámetros en el teorema de Bayes para realizar la predicción. Sin embargo, a diferencia de LDA, QDA asume que cada clase tiene su propia matriz de covarianza. Es decir, asume que una observación de la k-ésima clase es de la forma $X \sim N(\mu_k, \Sigma_k)$, donde Σ_k es una matriz de covarianza para la k-ésima clase. Bajo esta suposición, el clasificador de Bayes asigna una observación $X = x$ a la clase para la cual

$$\begin{aligned}\delta_k(x) &= -\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) - \frac{1}{2}\log|\Sigma_k| + \log(\pi_k) \\ &= -\frac{1}{2}x^T\Sigma_k^{-1}x + x^T\Sigma_k^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma_k^{-1}\mu_k - \frac{1}{2}\log|\Sigma_k| + \log(\pi_k) \qquad (3.29)\end{aligned}$$

es mayor. Entonces, el clasificador QDA implica insertar estimaciones para Σ_k, μ_k, y π_k en (3.29), y luego asignar una observación $X = x$ a la clase para la cual esta cantidad es mayor. A diferencia de (3.24), la cantidad x aparece como una función cuadrática en (3.29). Aquí es donde QDA obtiene su nombre.

Elegir LDA o QDA radica en el equilibrio entre sesgo y varianza. Cuando hay p predictores, estimar una matriz de covarianza requiere estimar $p(p+1)/2$ parámetros. QDA estima una matriz de covarianza separada para cada clase, para un total de $Kp(p+1)/2$ parámetros. Con 50 predictores, esto es un múltiplo de 1225, lo cual es una gran cantidad de parámetros. Al asumir en su lugar que las K clases comparten una matriz de covarianza común, el modelo LDA se vuelve lineal en x, lo que significa que hay Kp coeficientes lineales para estimar. En consecuencia, LDA es un clasificador mucho menos flexible que QDA, y por lo tanto tiene una varianza sustancialmente menor. Esto puede llevar potencialmente a un mejor rendimiento de predicción. Pero hay un compromiso: si la suposición de LDA de que las K clases comparten una matriz de covarianza común es incorrecta, entonces LDA puede sufrir de alto sesgo. En términos generales, LDA tiende a ser una mejor opción que QDA si hay relativamente pocas observaciones de entrenamiento y por lo tanto reducir la varianza es crucial. En contraste, se recomienda QDA si el conjunto de entrenamiento es muy grande, de modo que la varianza del clasificador no sea una preocupación importante, o si la suposición de una matriz de covarianza común para las K clases es claramente insostenible.

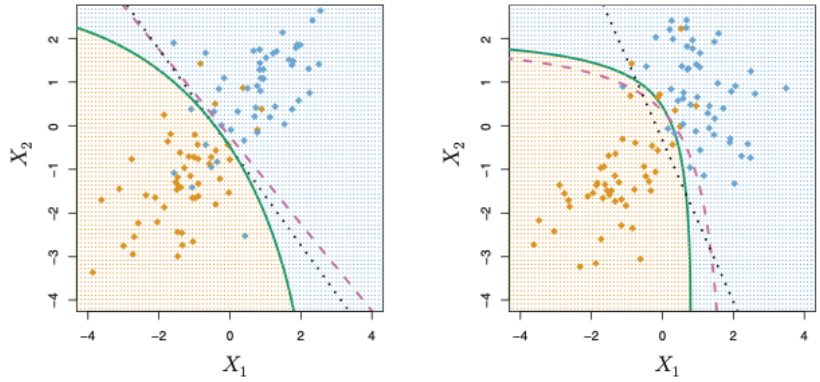

FIGURE XV: Izquierda: Las fronteras de decisión de Bayes (línea discontinua púrpura), LDA (línea punteada negra) y QDA (línea sólida verde) para un problema de dos clases con $\Sigma_1 = \Sigma_2$. El sombreado indica la regla de decisión de QDA. Dado que la frontera de decisión de Bayes es lineal, es más precisamente aproximada por LDA que por QDA. Derecha: Los detalles son los mismos que en el panel izquierdo, excepto que $\Sigma_1 \neq \Sigma_2$. Dado que la frontera de decisión de Bayes es no lineal, es más precisamente aproximada por QDA que por LDA.

La figura XV ilustra el rendimiento de LDA y QDA en dos escenarios. En el panel izquierdo, las dos clases gaussianas tienen una correlación común de 0.7 entre X_1 y X_2. Como resultado, la frontera de decisión de Bayes es lineal y es aproximada con precisión por la frontera de decisión de LDA. La frontera de decisión de QDA es inferior, porque sufre de mayor varianza sin una disminución correspondiente en el sesgo. En contraste, el panel derecho muestra una situación en la que la clase naranja tiene una correlación de 0.7 entre las variables y la clase azul tiene una correlación de -0.7. Ahora la frontera de decisión de Bayes es cuadrática, y por lo tanto QDA aproxima más precisamente esta frontera que LDA.

4 Evaluación y selección de modelos

Lo mas importante del curso.

- Empezamos por regresion. Tenemos una variable de salida Y que toma valores en un ontinuo y con p predictores-. Modelamos con una funcion y añadimos el temrino de error. Siempre tendremos este error. nosotros determinamos la $\hat{Y}$ con $\hat{f}$. El error cuadratico medio es la media de los errores al cuadrado.
- Como evaluamos el error de entrenamineto, que metricas, etc. En ridge, el parametro es beta, loque aprende el modelo. El coeficiente de regularizacion, es decir, λ, es el hiperparámetro, se lo damos nosotros. No free lunch teorem. algo muy importante!
- Empezamos a evaluar la calidad del ajuste. No dice nada relevante. R^2 no le gusta mucho a el. Esta metrica primero calcula el eror que obtendria el sistema de aprendizake que se nos puede ocurrir: dar como salida la media de todos nuestros ejemplos (TSS da como salida la media). Luego calcula la diferencia entre eso y la suma de los cuadrados y normalizandolo. HAciendolo peor que la media esta metrica daria negativa incluso (muy dificil). Gusta porque el error se acota entre 0 (dificil bajarla) y 1. No le gusta a el porque es engañoso, los problemas no son comparables y por tanto los valores de la metrica tampoco. Se suele usar el MSE porque no depende del numero de ejemplos.

  menor MSE train no garantiza menor MSE test! Interesamos en el modelo con mejor error de test.
- Diapositiva importante !! Sobreaprendizaje y subaprendizaje. EL modleo verde es el de menor MSE de train pero el azul el de menor MSE de test. Flexibilidad = complejidad de modelo, mejor se va a adaptar a los datos del modelo. Error de train y test alto = subaprendizaje, modelo demasiado simple para nuestros datos. Error de train bajo y test alto = sobreaprendizaje, modelo demasiado complejo. Al hacer exploracion hay que intentar conseguir siempre la curva para llegar al subaprendizaje y sobreaprendizaje, asi aseguramos que barremos todo el rango de hiperparametros.
- Tanto el amarillo como el azul son mas o menos buenos. A igualdad, preferimos modelos sencillos, ya que tiene mayor capacidad de generalizacion.
- Habria que explorar mas modelos, se sobreajuste
- Dos formas de estimar el error de test (1) teoricamente, cogemos un subconjunto datos y calculamos el error de test sobre otro conjunto de entrenamiento. (2) lo mimos ??? La varianza nos dice cuanto cambia la funcion si cambiamos algunos de los datos de entrenamiento. Modelos sencillos tienen poca varianza, variar un dato no lo cambia mucho. Modelos complejos tienen mucha varianza, variar un dato lo cambia mucho.
- Bias: def. el modelo amarillo de la 2.9 tiene mucho bias, modelo verde casi nulo.
- La curva roja es el error total, es decir, suma el bias, avarianza y el error irreducible.Minimos cuadrados es un modelo sobreaprendido. La flexibilidad de ridge y laso va entonces con $1/\lambda$, a mayor lambda menor lambda el modelo tiende al sobreaprendizaje.

- Ahora lo mismo peroapra clasificacion. PAra medir la calidad de modelo se suele usar el error de clasificacion: contamos el numero de ejemplos en los que nos estamos equivocando. No profundizara en el clasificador de bayes. LDA hace una aproximacion al clasificador de bayes porque no se conocen las probabilidades; LDA asume una distribucion normal.
- Como tenemos el ejemplo, si podemos usar el de Bayes
- Asumimos un problema de clasificacion binario. Ahi podemos construir una matriz de confusion. N^* son los negativos que estima el modelo ($N \neq N^*$).
- La curva ROC se usa para medir la calidad de clasificadores binarios

4.1 Selección de subconjuntos

4.1.1 Selección del mejor subconjunto

Para hacer la selección del mejor subconjunto, se debe ajustar una regresión de mínimos cuadrados distinta para cada combinación de los p predictores. Esto es, se ajustan todos los p modelos que contienen exactamente un predictor, los $\binom{p}{2} = p(p-1)/2$ modelos que contienen exactamente dos predictores, y así sucesivamente. Luego, se selecciona el mejor modelo.

El problema viene en elegir el mejor de entre las 2^p posibilidades consideradas. Esto se suele hacer en dos etapas:

i. Sea $\mathcal{M}_0$ el modelo nulo que no contiene ningún predictor. Este modelo predice la media de la muestra para cada observación.

ii. Para $k = 1, 2, \ldots, p$:

   (a) Ajustar todos los $\binom{p}{k}$ modelos que contienen exactamente k predictores.

   (b) Elegir el mejor modelo entre los $\binom{p}{k}$ modelos, y llamarlo $\mathcal{M}_k$. Aquí, "mejor" se refiere a tener el menor RSS o, equivalentemente, el mayor R^2. Tras esto, el problema se reduce de 2^p posibilidades a $p + 1$.

iii. Elegir un único "mejor" modelo de entre $\mathcal{M}_0, \ldots, \mathcal{M}_p$ usando predicción de error validada de forma cruzada, C_p (AIC), BIC, o R^2 ajustado.

Para elegir el mejor modelo hay que elegir entre los $p + 1$ modelos $\mathcal{M}_i$, con $i = 0, \ldots, p$. Hay que tener en cuenta que el RSS de estos modelos decrece de forma monótona, mientras que el R^2 aumenta de forma monótona. Por tanto, si se usa estos estadísticos para elegir el mejor modelo, siempre se acabará con un modelo que incluya todas las variables. El problema es que un RSS bajo o un R^2 alto indica un modelo con un error de entrenamiento bajo, mientras que lo que se quiere es elegir un modelo con un error de *test* bajo. Por tanto, en el paso 3, se usa la predicción de error validada de forma cruzada, C_p, BIC o R^2 ajustado para elegir entre $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p$.

4.1.2 Selección por pasos

Por motivos computacionales, la selección del mejor subconjunto no sirve para p grandes, caso donde puede sufrir de porblemas estadísticos. Cuando mayor sea el espacio de búsqueda, mayor será la posibilidad de encontrar modelos que ajusten bien el conjunto de entrenamiento, aunque no tenga buen poder predictivo. Entonces, un gran espacio de búsqueda puede conducir a *overfitting* y una gran variación de los coeficientes estimados. Los modelos de selección por pasos exploran un conjunto restringido de modelos, por lo que resultan una buena alterantiva.

Selección por pasos hacia adelante

Este método resulta más eficiente computacionalmente que la selcción del mejor subconjunto. Este método comienza con un modelo que no contenga predictores, y va añadiendo predictores al modelo, uno a uno, hasta que todos los predictores están dentro del modelo. En particular, en cada paso, se añade la variable que dé la mayor mejora al ajuste. Formalmente:

i. Sea $\mathcal{M}_0$ el modelo nulo, que no contiene predictores.

ii. Para $k = 0, \ldots, p-1$:

   (a) Considera los $p-k$ modelos que aumentan los predictores en $\mathcal{M}_k$ con un predictor adicional.

   (b) Elige el mejor entre estos $p-k$ modelos y lo denota $\mathcal{M}_{k+1}$. Aquí el "mejor" es aquel con menor RSS o mayor R^2.

iii. Elige el mejor modelo entre $\mathcal{M}_0, \ldots, \mathcal{M}_p$ usando predicción de error validada de forma cruzada, C_p (AIC), (BIC) o R^2 ajustado.

A diferencia de la selección del mejor subconjunto, que necesita ajustar 2^p modelos, la selección por pasos hacia adelante necesita ajustar un modelo nulo, junto con $p-k$ modelos en la iteración k-ésima para $k = 0, \ldots, p-1$. Esto resulta en un total de $1 \sum_{k=0}^{p-1}(p-k) = 1+p(p+1)/2$ modelos.

En el segundo paso, el apartado (b), se debe elegir el mejor modelo entre los $p-k$ modelos que aumentan $\mathcal{M}_k$ con un predictor adicional. Esto se puede hacer eligiendo el modelo con menor RSS o mayor R^2. Sin embargo, en el paso 3, se debe elegir el mejor modelo entre un conjunto de modelos con diferente número de variables. Esto es más complicado y se discute en la sección 6.1.3.

La ventaja computacional del método de selección por pasos hacia adelante sobre la selección del mejor subconjunto es clara. Aunque el método de selección por pasos hacia adelante tiende a funcionar bien en la práctica, no está garantizado que encuentre el mejor modelo posible de entre los 2^p modelos que contienen subconjuntos de los p predictores. Por ejemplo, sea un conjunto de datos con $p = 3$ predictores, el mejor modelo de una variable contiene X_1, y el mejor modelo de dos variables contiene X_2 y X_3. Entonces, la selección por pasos hacia adelante no seleccionará el mejor modelo de dos variables, porque $\mathcal{M}_1$ contendrá X_1, por lo que $\mathcal{M}_2$ también debe contener X_1 junto con una variable adicional.

La selección por pasos hacia adelante se puede aplicar incluso en el caso de gran dimensión donde $n < p$, aunque en este caso, solo se pueden construir submodelos $M_0, \ldots, M_{n-1}$, ya que cada submodelo se ajusta utilizando mínimos cuadrados, lo que no dará una solución única si $p \geq n$.

Selección por pasos hacia atrás

Este método comienza con el modelo de mínimos cuadrados que contiene todos los predictores, y luego elimina uno a uno los predictores que menos contribuyen al ajuste. Formalmente:

i. Sea $\mathcal{M}_p$ el modelo completo que contiene los p predictores.

ii. Para $k = p, p-1, \ldots, 1$:

(a) Considera los k modelos que contienen todos menos uno de los predictores en $\mathcal{M}_k$, para un total de $k-1$ predictores.

(b) Elige el mejor entre estos k modelos y lo denota $\mathcal{M}_{k-1}$. Aquí el "mejor" es aquel con menor RSS o mayor R^2.

iii. Elige el mejor modelo entre $\mathcal{M}_0, \ldots, \mathcal{M}_p$ usando predicción de error validada de forma cruzada, C_p (AIC), (BIC) o R^2 ajustado.

La selección por pasos hacia atrás también necesita ajustar $1 + p(p+1)/2$ modelos, al igual que la selección por pasos hacia adelante, y también puede aplicarse para el caso en el que p es demasiado grande como para aplicar la selección del mejor conjunto. Este método tampoco garantiza encontrar el mejor modelo de entre los 2^p posibles.

La selección por pasos hacia detrás no se puede aplicar en el caso en el que $n < p$, ya que el modelo completo no se puede ajustar en este caso.

Modelos híbridos

En general, no se obtienen los mismos modelos de selección por pasos hacia adelante y hacia atrás. Como alternativa, se pueden considerar versiones híbridas de selección por pasos hacia adelante y hacia atrás, en las que las variables se agregan al modelo secuencialmente, de manera análoga a la selección hacia adelante, pero después de agregar cada nueva variable, el método puede eliminar cualquier variable que ya no proporcione una mejora en el ajuste del modelo. Este enfoque intenta imitar más de cerca la selección del mejor subconjunto, mientras mantiene las ventajas computacionales de la selección por pasos hacia adelante y hacia atrás.

4.1.3 Selección del modelo óptimo

$$C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2) \tag{4.1}$$

$$AIC = \frac{1}{n\hat{\sigma}^2}(RSS + 2d\hat{\sigma}^2) \tag{4.2}$$

$$BIC = \frac{1}{n}(RSS + \log(n)d\hat{\sigma}^2) \tag{4.3}$$

$$R^2_{\text{ajustado}} = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)} \tag{4.4}$$

Validación y validación cruzada

Reduciendo de error

En situacion de subaprendizaje, añadir mas datos no va a ayudar. PAra resolverlo hay que ir a un modelo más complejo. Podemos añadir caracteristicas o variables nuevas y/o decrementar la regularización (en ridge lasso seria disminuir λ) (este es el más directo y sencillo).

En situacion de sobreaprendizaje, añadir mas datos si que ayuda. Hay que ir a un modelo más simple. Podemos eliminar caracteristicas o variables y/o incrementar la regularización .

5 K vecinos más próximos

En teoría, siempre se desearía predecir respuestas cualitativas usando el clasificador de Bayes. Pero para datos reales, no se conoce la distribución condicional de Y dado X, por lo que calcular el clasificador de Bayes es imposible. Por lo tanto, el clasificador de Bayes sirve como un estándar inalcanzable contra el cual comparar otros métodos. Muchos enfoques intentan estimar la distribución condicional de Y dado X, y luego clasificar una observación dada a la clase con la mayor probabilidad estimada. Uno de estos métodos es el clasificador de K-vecinos más próximos (KNN). Sea un entero positivo K y una observación de prueba x_0, el clasificador KNN primero identifica los K puntos en los datos de entrenamiento que están más cerca de x_0, representados por $\mathcal{N}_0$. Luego estima la probabilidad condicional para la clase j como la fracción de puntos en $\mathcal{N}_0$ cuyos valores de respuesta son iguales a j:

$$\Pr(Y = j|X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = j) \tag{5.1}$$

Finalmente, KNN aplica la regla de Bayes y clasifica la observación de prueba x_0 a la clase con la mayor probabilidad.

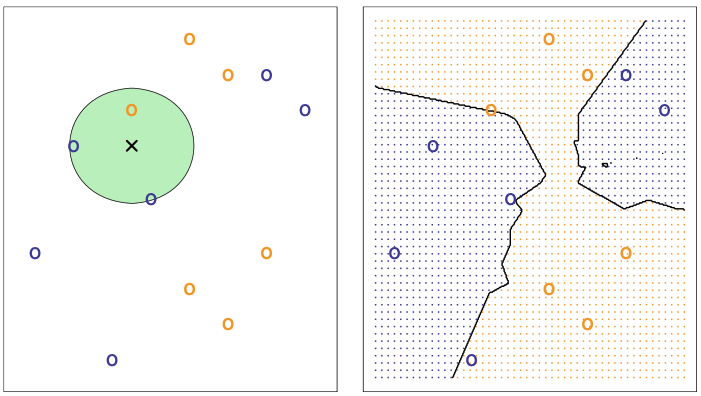

FIGURE XVI: El enfoque KNN, usando $K = 3$, se ilustra en una situación simple con seis observaciones azules y seis observaciones naranjas. Izquierda: una observación de prueba para la cual se desea una etiqueta de clase predicha se muestra como una cruz negra. Se identifican los tres puntos más cercanos a la observación de prueba, y se predice que la observación de prueba pertenece a la clase que ocurre con mayor frecuencia, en este caso azul. Derecha: La frontera de decisión KNN para este ejemplo se muestra en negro. La cuadrícula azul indica la región en la cual una observación de prueba será asignada a la clase azul, y la cuadrícula naranja indica la región en la cual será asignada a la clase naranja.

La figura XVI proporciona un ejemplo del enfoque KNN. En el panel izquierdo, se ha graficado un pequeño conjunto de datos de entrenamiento que consiste en seis observaciones azules y seis naranjas. El objetivo es hacer una predicción para el punto etiquetado con la

cruz negra. Supongamos que se elige $K = 3$. Entonces KNN primero identificará las tres observaciones que están más cerca de la cruz. Este vecindario se muestra como un círculo. Consiste en dos puntos azules y un punto naranja, resultando en probabilidades estimadas de $2/3$ para la clase azul y $1/3$ para la clase naranja. Por lo tanto, KNN predecirá que la cruz negra pertenece a la clase azul. En el panel derecho de la figura XVI se ha aplicado el enfoque KNN con $K = 3$ en todos los valores posibles para X_1 y X_2, y se ha dibujado la correspondiente frontera de decisión KNN. A pesar de que es un enfoque muy simple, KNN puede producir clasificadores que están sorprendentemente cerca del clasificador de Bayes óptimo. La figura XVII muestra la frontera de decisión KNN, usando $K = 10$, cuando se aplica al conjunto de datos simulado más grande de la figura X. Nótese que aunque el clasificador KNN no conoce la distribución verdadera, la frontera de decisión KNN está muy cerca de la del clasificador de Bayes. La tasa de error de prueba usando KNN es 0.1363, que está cerca de la tasa de error de Bayes de 0.1304.

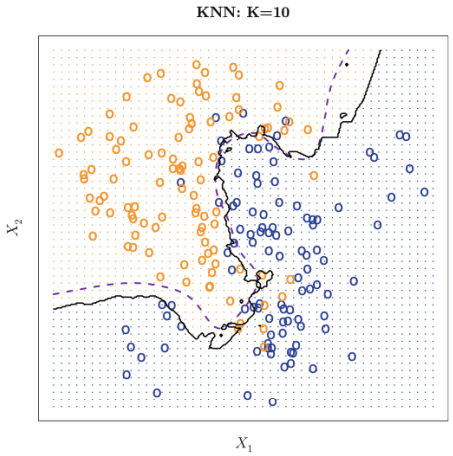

FIGURE XVII: La curva negra indica la frontera de decisión KNN en los datos de la figura X, usando $K = 10$. La frontera de decisión de Bayes se muestra como una línea discontinua púrpura. Las fronteras de decisión KNN y de Bayes son muy similares.

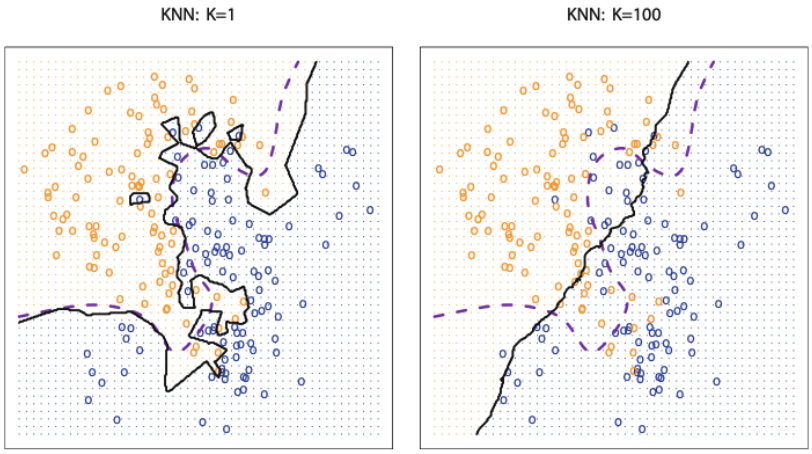

FIGURE XVIII: Una comparación de las fronteras de decisión KNN (curvas negras sólidas) obtenidas usando $K = 1$ y $K = 100$ en los datos de la figura X. Con $K = 1$, la frontera de decisión es excesivamente flexible, mientras que con $K = 100$ no es suficientemente flexible. La frontera de decisión de Bayes se muestra como una línea discontinua púrpura.

La elección de K tiene un efecto drástico en el clasificador KNN obtenido. La figura XVIII muestra dos ajustes KNN a los datos simulados de la figura X, usando $K = 1$ y $K = 100$. Cuando $K = 1$, la frontera de decisión es excesivamente flexible y encuentra patrones en los datos que no corresponden a la frontera de decisión de Bayes. Esto corresponde a un clasificador que tiene bajo sesgo pero muy alta varianza. A medida que K crece, el método se vuelve menos flexible y produce una frontera de decisión que es casi lineal. Esto corresponde a un clasificador de baja varianza pero alto sesgo. En este conjunto de datos simulado, ni $K = 1$ ni $K = 100$ dan buenas predicciones: tienen tasas de error de prueba de 0.1695 y 0.1925, respectivamente.

Al igual que en el contexto de regresión, no hay una relación estrecha entre la tasa de error de entrenamiento y la tasa de error de prueba. Con $K = 1$, la tasa de error de entrenamiento de KNN es 0, pero la tasa de error de prueba puede ser bastante alta. En general, a medida que se usan métodos de clasificación más flexibles, la tasa de error de entrenamiento disminuirá pero la tasa de error de prueba puede no hacerlo. En la figura XIX se han graficado los errores de prueba y de entrenamiento de KNN como una función de $1/K$. A medida que $1/K$ aumenta, el método se vuelve más flexible. Como en el contexto de regresión, la tasa de error de entrenamiento disminuye consistentemente a medida que aumenta la flexibilidad. Sin embargo, el error de prueba exhibe una forma característica de U, disminuyendo al principio (con un mínimo en aproximadamente $K = 10$) antes de aumentar nuevamente cuando el método se vuelve excesivamente flexible y sobreajusta.

En ambos contextos, de regresión y clasificación, elegir el nivel correcto de flexibilidad es crítico para el éxito de cualquier método de aprendizaje estadístico. El equilibrio entre sesgo y varianza, y la resultante forma de U en el error de prueba, pueden hacer de esta una tarea difícil.

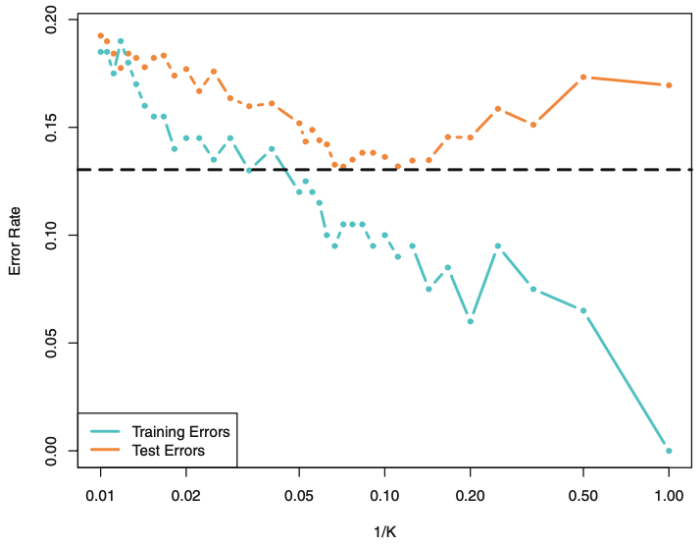

FIGURE XIX: La tasa de error de entrenamiento KNN (azul, 200 observaciones) y la tasa de error de prueba (naranja, 5,000 observaciones) en los datos de la figura X, a medida que el nivel de flexibilidad (evaluado usando $1/K$) aumenta, o equivalentemente, a medida que el número de vecinos K disminuye. La línea discontinua negra indica la tasa de error de Bayes. La irregularidad de las curvas se debe al pequeño tamaño del conjunto de datos de entrenamiento.

6 KNN

Si tenemos un piso de 85 m2 y 0.4 km de distancia, intuitivamente se parece más al primer piso que al segundo. Pero si calculamos la distancia euclidiana, obtenemos que la distancia al primer piso esta mucho mas lejos que el segundo. Sin embargo, si cambiamos la escala de km a m, esto no sucede. Por esto hay que estandarizar:

$$x \to \frac{x - \bar{x}}{\text{std}x} = x' \tag{6.1}$$

Asi, la media de x' es 0 y la desviacion estandar 1. Para dar la salida hay que desentandarizar !!!

Numero de vecino es el unico hiperparametro. Hay que estandarizar de forma obligatoria, pocos vecinos, sobreaprendizaje, muchos vecinos, subaprendizaje.

Pcos datos de entrenamiento y alta dimensionalidad de entrada es el peor escenario en aprendizaje estadistico, no le sienta bien a ningun algoritmo.

	superficie (m2)	distancia (km)
superficie	70	0.1
distancia	90	5.0

7 Arboles

En un arbol de decision se tiene muy claro de donde viene la salida (interpretable), pero no puede competir con los mejores modelos de predicción. Solo veremos CART, que vale para regresión y clasificación.

EXAMPLE:

Los menores de 4.5 (estricto) se da el valor directametne. Si es mayor igual de 4.5, analizamos los golpes, si es menos, 6, si es mayor, 6.74. El primer nodo de años divide el eje de años en dos partes (no tienen por qué ser iguales). En el siguiente nodo se divide la parte correspondiente a una de las partes anteriores por el eje Hits. Divides en regiones y cuando una muestra cae en una región, se da la media de los datos de entrenamiento de esa región. La division en años, puede hacerse de tantas formas como años se disponga en los datos.

REGRESSION TREES: el tamaño del arbol es el hiperparametro que controlara el subaprendizaje o subaprendizaje del algoritmo (determina el tamaño del arbol). arbol pequeño subapredndio. Buscamos la division de regiones que minimiza la formula del RSS vista.

USAMOS METODOS APROXIMADO. RECURSIVE BINARY SPLITTING.

En cada punto escogemos la mejor variable y el mejor umbral. Decisiones individuales y no globales para que sea computacionalmente mas eficiente. No hay garantia de encontrar el optimo para si una bueno. El criterio es la formaulas que salen ahi para R1 y R2 minimizando ese error.

Ejemplo 7.0.1. *Sea dos variables* $X_1 = (-3, -2, 0, 2, 5)$ *e* $Y = (4, 6, -2, 8, 10)$. *Buscamos las posibles particiones para esas variables. En este caso entre los datos (punto medio),* $(-2.5, -1, 1, 3.5)$. *Cada umbral divide el conjunto en dos regiones, la de la izquierda de los datos y la de la derecha. Habría que calcular los errores, ejemplo con particion en +1.*

Esta particion nos deja dos regiones $R_1 = \{(-3, 4), (-2, 6), (0, -2)\}$ *y* $R_2 = \{(2, 8), (5, 10)\}$. *Calculamos la media de las salidas de cada región:* $\hat{y}_{R_1} = 8/3$ *y* $\hat{y}_{R_2} = 9$. *Calculamos ahora el error*

$$\underbrace{(4 - 8/3)^2 + (6 - 8/3)^2 + (-2 - 8/3)^2}_{R_1} + \underbrace{(8 - 9)^2 + (10 - 9)^2}_{R_2} \tag{7.1}$$

Lo repetimos para los 4 umbrales posibles y nos quedamos con el que minimice el error. Luego repetimos con X_2 *y nos quedamos con el que minimice el error. Así con todas las variables y umbrales posibles.*

Paramos de dividir en una rama cuando se cumple el criterio de parada (esto nos dara por tanto el tamaño del arbol) usamos el tamaño minimo de nodo: cuando en un nodo hay menos de los ejemplos fijados, dejamos de dividirlo. Para sobreaprender, el tamaño de nodo deberá ser pequeño, por lo que el arbol sera grande. Este es el que mejor funciona en general. Unico hiperparametro que usaremos.

Hay otro criterio para parar: que todas las salidas de un nodo sean las mismas, no tendria sentido seguir dividiendo (este se ve mejor en clasificacion).

Problema de maxima profundidad, todas las ramas u hojas son del mimso tamaño

Otro se calcula el error de la division y si no mejora en un porcentaje, se para (comparar el error de una region con el error de la division de ambas (sumado obvio con la formula)). Problema: hacer mas divisiones puede mejorar el error, aunque una concreta no lo haga.

De cada nodo tenemos media y desviacion, por lo que podemos dar confianza ($\pm$) en la prediccion.

PODADO DEL ARBOL (PRUNING):

Probar quitar cada uno de los nodos es ineficiente. Usamos el cost complexity pruning. Podamos los nodos que nos aumenten menos el error, (de forma local, no global). EL TAMAÑO DEL ARBOL ES EL NUMERO DE REGIONES (NUMERO DE HOJAS n) Construimos el arbol de n-1 hojas y podamos el nodo que menos aumente el error (NO PODEMOS PODAR HOJAS O COMBINACIONES QUE NO VIENEN DEL MISMO NODO INMEDIATO) (en ejemplo, 5.487 y 4.622 no son una rama que podar). El que gana se une, dando un arbol de n-1 hojas. Repetimos el proceso hasta que no podamos mas.lo que da la grafica que vemos. En esa grafica vemos que el de 2 esta bastante bien. El profe prefiere no podar y usar el tamaño minimo de nodo.

EN CLASIFICACION.

Lo mismo pero sin usar el error cuadratico como error. La salida ahora sera la categoria mayoritaria en la region donde caiga el dato a predecir. Lo intuitivo sería usar el error de clasificacion. Esto es para binary splitting cuidado! Usamos el indice de Gini o la entropia.

- El indice de Ginny (mide la varianza entre todas las clases)
- Entropia: probabilidad por logaritmo de la probabilida y sumado sobre todas las clases. (- para que a probabilidad pequeña de valor grande). Usamos este criterio para hacer el proceso de recursive binary spliting (hacer crecer el arbol)

En la figura, los puntos de principio y fin de arco, es decir, 0 y 1, son los nodos puros ya que solo hay una clase. El maximo es el error en 0.5. El del error de clasificacion no aumenta lo suficientemente rapido para que sea un buen criterio.

El numero de muestras de una region influye en el alculo del error de regresión. En clasificación, no.

Ejemplo 7.0.2.

$$R_1 : \ 25\ 1 \ : \ -0\log(0) + 1\log(1) = 0 \tag{7.2}$$

$$R_1 : \ 25\ 1 \ y \ 50\ 0 \ : \ -\frac{2}{3}\log\left(\frac{2}{3}\right) - \frac{1}{3}\log\left(\frac{1}{3}\right) = 0.918 \tag{7.3}$$

Para otro problema distinto, donde

$$R_1 : \ 50\ 1 \ : \ -0\log(0) - 1\log(1) \tag{7.4}$$

$$R_1 : \ 10\ 1 \ y \ 20\ 0 \ : \ -\frac{2}{3}\log\left(\frac{2}{3}\right) - \frac{1}{3}\log\left(\frac{1}{3}\right) \tag{7.5}$$

Con criterios quedan iguales. El mejor sin embargo es el segundo, ya que clasifico 50 bien y luego en R_2 me equivoco en 10.

Por esto, no usamos tal cual entropia, sino

$$\frac{|R_1|}{|R|}S_1 + \frac{|R_2|}{|R|}S_2 \tag{7.6}$$

donde $|R|$ es el tamaño del nodo R. nodos puros cuanto mas grandes mejor.

Ejemplo 7.0.3. *Usamos el recursive binary tree. Sea una variable* $X_1 = (-4, -2, -2, -1, 1, 3)$ *y* $Y = (-, -, +, +, -, +)$. *Posibles puntos de división:* $(-3, -1.5, 0, 2)$. *Hay que coger todos los umbrales y quedarnos con el menor como antes. Cogemos por ejemplo el* -1.5. *Tamaño de* R_1 *es 3, tamaño de R es 6,* $p_+ = 2/3$ *y* $p_- = 1/2$. *Tamaño de* R_2 *es 3,* $p_+ = 1/3$ *y* $p_- = 2/3$. *Calculamos la entropia de cada región*

$$S = \underbrace{\frac{3}{6}\left(-\frac{2}{3}\log\left(\frac{2}{3}\right) - \frac{1}{3}\log\left(\frac{1}{3}\right)\right)}_{R_1} + \underbrace{\frac{3}{6}\left(-\frac{1}{3}\log\left(\frac{1}{3}\right) - -\frac{2}{3}\log\left(\frac{2}{3}\right)\right)}_{R_2} \tag{7.7}$$

$$\tag{7.8}$$

Repetimos para todos los umbrales y nos quedamos con el que minimice el error. Luego repetimos con cada variable y nos quedamos con el que minimice el error.

ESTE ES EL UNICO DE MODELOS QUE VAMOS A VER QUE TRABAJA CON VARIABLES CUALITATIVAS DE FORMA NATURAL.

PREDICTORES CATEGORICAS:

Ejemplo 7.0.4. *Sea una variable con cuatro posibles valores (*$q = 4$*)* $X_1 = (A, B, C, D)$, *para particionar podemos hacer del 7 formas distintas:*

$$\{(A, BCD), (B, ACD), (C, ABD), (D, ABC), (AB, CD), (AC, BD), (AD, BC)\} \tag{7.9}$$

Podemos reducir el numero de particiones a calcular Sea $X_1 \in (A, B, C, D)$. *Supongamos que de los ejemplos de* $X_1 = A$, $p_+ = 0.2$ *y* $p_- = 0.8$. *Para* $X_1 = B$, $p_+ = 0.3$ *y* $p_- = 0.7$. *Para* $X_1 = C$, $p_+ = 0.4$ *y* $p_- = 0.6$. *Para* $X_1 = D$, $p_+ = 0.1$ *y* $p_- = 0.9$.

Ordenamos en orden creciente de la clase positiva. El valor más bajo es D con 0.1, el siguiente A con 0.2, luego B con 0.3 y finalmente C con 0.4. Analizando umbrales ahora (que serían tres), garantizamos solución optima en cuanto a gini y entropia. ESTO SOLO ES VALIDO PARA CLASIFICACION BINARIA, independientemente de q. PARA MULTICLASIFICACION hay que probar todas las posibles particiones sin hacer este truco de reducir.

Este truco se puede usar para regresión pero de la sigueinte forma. En vez de usar las p_+, calculamos la media de los ejemplos apra cada uno de los posibles valores de la clase X_1. Ordenamos por medias de forma creciente y calculamos los umbrales. Aplicamos lo de siempre.