Métodos de ensemble aplicados a la predicción de la cantidad energía eólica obtenida

Trabajo de Fin de Grado

Grado en Matemáticas

Autor: José Benjumeda Rubio

Tutores: Jose Luis Torrecilla Noguerales y Luis Alberto Rodríguez Ramírez

Curso 2020-2021

1 Introducción

La estadística y el análisis de datos permiten llegar cada vez más lejos en el desarrollo de inteligencia artificial, buscando combinar el razonamiento humano con la velocidad y capacidad de procesamiento de una máquina. En este trabajo se desarrollará un método de ensemble que aprenderá la relación entre un conjunto de datos conocidos para obtener otros desconocidos, es decir, se estudiará el problema de regresión.

Tras explicar el contexto estadístico sobre el que se define el problema de regresión, se explicarán los modelos que compondrán el método de ensemble: el perceptrón simple y la *Support Vector Machine*, que implementan de dos maneras distintas una poderosa idea: la transformación no lineal adecuada permite obtener a partir de unos datos que no pueden explicarse de manera lineal, otros que sí se pueden aproximar satisfactoriamente mediante un modelo lineal.

Para dar apoyo a las explicaciones teóricas sobre estos modelos, se entrenarán para aprender la relación entre la cantidad de energía eólica obtenida cada hora de cada día en el parque eólico experimental de Sotavento, en Galicia, y las predicciones de los valores de módulo y dirección del vector de viento medido tanto a 10 como a 100 metros de altura, la temperatura medida a 2 metros de altura y la presión en la superficie.

1.1 ¿Qué es un problema de regresión?

Un *problema de regresión* consiste en asignar a una nueva observación de una variable aleatoria un valor numérico. Esto nos lleva a nuestra primera definición formal.

Definición 1.1. Dado un espacio de probabilidad $(\Omega, \mathcal{A}, P)$, y un espacio medible (S, Σ), una *variable aleatoria* X es una función

$$X : \Omega \to S$$

que es $\mathcal{A}/\Sigma$-medible. Si S es $\mathbb{R}$ y Σ es $\mathcal{B}(\mathbb{R})$, X es una variable aleatoria real.

En un problema de regresión hay dos o más variables: una variable dependiente Y, y n variables explicativas o independientes $X_i : i = 1, \ldots, n$. Un problema de regresión simple es aquel en el que hay una única variable independiente, y un problema de regresión múltiple es uno en el que hay al menos dos variables independientes. Además, si suponemos una relación lineal entre la variable dependiente y las independientes, estaremos haciendo regresión lineal, y, si no, regresión no lineal.

El tipo de regresión a utilizar viene determinado por la relación que haya entre los datos: si un porcentaje alto de ésta es lineal, obtendremos buenos resultados con un modelo lineal, pero en casos en los que la relación sea altamente no lineal, un modelo lineal no producirá buenos resultados por sí solo.

Parece que lo más intuitivo es, en este último caso, intentar ajustar los datos con una función no lineal. Sin embargo, obtendremos mucho mejores resultados transformando los datos de manera no lineal para luego poder utilizar regresión lineal. De esta idea parten las redes neuronales y las *support vector machine* que veremos más adelante, pero primero vamos a profundizar algo más en la regresión lineal.

1.1.1 Regresión lineal simple

En un problema de regresión lineal simple se busca explicar una variable aleatoria Y a partir de otra, X, asumiendo una relación lineal entre ambas más un error que sigue una distribución normal de media 0:

$$Y = f(X) = \beta_0 + \beta_1 X + \epsilon,$$

donde β_0 es el corte de la recta con el eje OY y β_1 es la pendiente de la recta, y $f(X)$ es el modelo lineal. Que exista este ϵ significa que la variable aleatoria X explica un cierto porcentaje de Y, que suele no ser su totalidad.

En el contexto habitual de un problema de regresión lineal, no conocemos los valores de β_0 ni de β_1, así que los estimamos mediante $\hat{\beta}_0$ y $\hat{\beta}_1$. La manera de estimarlos es minimizando el error cuadrático medio visto como función de $\hat{\beta}_0$ y $\hat{\beta}_1$, a partir de una muestra de tamaño n $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$:

$$J(\hat{\beta}_0, \hat{\beta}_1) = \frac{1}{n}\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 \tag{1}$$

Estudiando la derivada, vemos que este mínimo siempre existirá:

$$\begin{aligned} J(\hat{\beta}_0, \hat{\beta}_1) &= \frac{1}{n}\sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 = \\ &= \frac{1}{n}\sum_{i=1}^{n}(y_i^2 + \hat{\beta}_0^2 + \hat{\beta}_1^2 x_i^2 + 2\hat{\beta}_0\hat{\beta}_1 x_i - 2\hat{\beta}_0 y_i - 2\hat{\beta}_1 y_i x_i) = \\ &= \overline{y^2} + \hat{\beta}_0^2 + \hat{\beta}_1^2\overline{x^2} + 2\hat{\beta}_0\hat{\beta}_1\overline{x} - 2\hat{\beta}_0\overline{y} - 2\hat{\beta}_1\overline{xy} = \\ &= (\hat{\beta}_0, \hat{\beta}_1)\begin{pmatrix} 1 & \overline{x} \\ \overline{x} & \overline{x^2} \end{pmatrix}\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} - 2\overline{y}\hat{\beta}_0 - 2\overline{xy}\hat{\beta}_1 + \overline{y^2} \end{aligned} \tag{2}$$

El primer término,

$$(\hat{\beta}_0, \hat{\beta}_1) \begin{pmatrix} 1 & \overline{x} \\ \overline{x} & \overline{x^2} \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix}, \tag{3}$$

es una forma cuadrática definida positiva, pues su determinante es la varianza de x, que es la suma de los cuadrados de las diferencias de cada muestra x con la media $\overline{x}$, y, por definición, cualquier suma de cuadrados de números reales es positiva. Por lo tanto, por ser una forma cuadrática definida positiva, tiene un mínimo.

Los otros dos términos lineales y el término constante no cambian este hecho, ya que al derivar quedan solo constantes, y sumando constantes no cambiamos la cantidad de soluciones de una ecuación polinómica.

Derivamos respecto de β_0 y de β_1 y obtenemos los valores en que hacen mínimo el error cuadrático:

$$\frac{\partial}{\partial \hat{\beta}_0} J(\hat{\beta}_0, \hat{\beta}_1) = 2\beta_0 + 2\overline{x}\hat{\beta}_1 - 2\overline{y} = 0$$

$$\hat{\beta}_0 = \overline{y} - \overline{x}\frac{COV(x,y)}{\sigma^2(x)} \tag{4}$$

$$\frac{\partial}{\partial \hat{\beta}_1} J(\hat{\beta}_0, \hat{\beta}_1) = 2\beta_0\overline{x} + 2\overline{x^2}\hat{\beta}_1 - 2\overline{xy} = 0$$

$$\hat{\beta}_1 = \frac{\overline{xy} - \overline{x}\ \overline{y}}{\overline{x^2} - \overline{x}^2} = \frac{COV(x,y)}{\sigma^2(x)} \tag{5}$$

En las ecuaciones de las derivadas parciales, hemos estimado la media poblacional con la media muestral. Este estimador es insesgado:

$$E[\overline{x}] = E[\frac{1}{n}\sum_{i=1}^{n} x_i] = \frac{1}{n}\sum_{i=1}^{n} E[x_i] = \frac{1}{n} nE[x] = E[x] \tag{6}$$

Como los estimadores utilizados son combinaciones de estimadores insesgados, también son insesgados. Nada más hay que pensar que en media, los estimadores insesgados que lo componen pueden considerarse valores poblacionales en lugar de muestrales, con lo que la combinación también será un valor poblacional.

No solo nos interesa saber que en media, nuestros estimadores coinciden con los valores poblacionales, sino cómo de lejos del valor poblacional estará una estimación cualquiera. Para esto miramos la varianza, que tiene la siguiente fórmula:

$$\sigma^2(\hat{\beta}_0) = \sigma^2(\epsilon)\left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right) \qquad \sigma^2(\hat{\beta}_1) = \frac{\sigma^2(\epsilon)}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Ambas fórmulas tienen la condición de que el error ϵ y la covarianza sean incorrelados, aunque incluso si esto no ocurre es probable que obtengamos buenas aproximaciones utilizándolas.

De ellas podemos destacar que al estar la varianza de x en el denominador, esto nos beneficia. Cuanto más extendidas estén las x, mejor estimaremos los parámetros. Podemos imaginar que queremos trazar una recta uniendo dos puntos: cuando estos están muy cerca, un pequeño error lleva a una recta muy diferente de la correcta, y en cambio, si están lejos, para que la recta varíe tanto necesitaríamos haber cometido un error mucho más grande.

Normalmente, ϵ es desconocido,y se estima a partir de la muestra utilizando el error cuadrático J para β_0 y β_1 fijos:

$$\hat{\epsilon} = \sqrt{\frac{J(\hat{\beta}_0, \hat{\beta}_1)}{n-2}}$$

A partir de la varianza de los estimadores obtenemos un intervalo de confianza del 95% muy sencillo:

Para β_0 : $\quad \hat{\beta}_0 \pm 2\sigma(\hat{\beta}_0)$ $\qquad\qquad$ Para β_1 : $\quad \hat{\beta}_1 \pm 2\sigma(\hat{\beta}_1)$

Los errores cuadráticos también son útiles a la hora de hacer contraste de hipótesis, que consiste en formular una hipótesis o afirmación sobre los datos, y comprobar si estos aportan, una vez fijado cierto nivel de confianza, suficiente evidencia como para probar que la afirmación es falsa, es decir, rechazar H_0 y aceptar la contraria, H_1.

Un contraste de hipótesis muy frecuente es afirmar que no existe ninguna relación entre X e Y. Matemáticamente:

$$\begin{aligned} H_0 &\equiv \beta_1 = 0 \\ H_1 &\equiv \beta_1 \neq 0 \end{aligned}$$

Para comprobar la hipótesis alternativa, tenemos que asegurarnos de que β_1 está lo suficientemente lejos de 0 como para estar seguros de que no ha sido fruto del azar. Para esto, suponemos que β_1 sí que es 0, y damos un intervalo de confianza α en el que obtendremos nuestra estimación si éste es el caso. Si la estimación que obtenemos a partir de la muestra no pertenece al intervalo, entonces rechazamos.

El tamaño del intervalo de confianza está directamente relacionado con la varianza de nuestro estimador. Si la varianza es alta, es probable que teniendo, por ejemplo, $\beta_1 = 0$, obtengamos una estimación $\hat{\beta}_1 >> 0$, con lo cual, será más difícil rechazar la hipótesis, y el intervalo de confianza α será mayor. En cambio, si tenemos una varianza muy cercana a cero, muestras que nos proporcionen

estimaciones también cercanas a cero pueden ser suficiente para descartar $\hat{\beta}_1 = 0$.

Utilizaremos el estadístico

$$t = \frac{\hat{\beta}_1 - 0}{\sigma(\hat{\beta}_1)}, \tag{7}$$

para ver cuántas $\sigma(\hat{\beta}_1)$ dista $\hat{\beta}_1$ de 0. Si verdaderamente no hay ninguna relación entre X e Y, es decir, si $\hat{\beta}_1$ es 0, esperamos tener una distribución t-student, con n-2 grados de libertad, que significa que lo más probable es obtener valores muy cercanos a 0 o el propio 0, sin importar si son algo mayores o algo menores (esta distribución es simétrica), y los que se vayan alejando más serán mucho menos probables. A partir de ± 30, la t-student y la normal(0,1) son muy parecidas.

Como tenemos una distribución fija conocida, podemos calcular la probabilidad de cada valor obtenido. Si, por ejemplo, el intervalo de confianza es del 95% y obtenemos un valor de $\hat{\beta}_1$ que pertenezca a $t_{\alpha/2}$ o a $t_{1-\alpha/2} = -t_{\alpha/2}$, descartamos la hipótesis.

1.1.2 Regresión lineal múltiple

La regresión lineal múltiple contempla el problema en el que hay que predecir el valor de una variable dependiente a partir de n variables independientes.

Una primera aproximación a este problema sería hacer n regresiones lineales independientes, pero habría que decidir qué valor de Y le corresponde a cada muestra, pues cada modelo nos daría un valor y acabaríamos teniendo n valores. Además, cada predicción se basaría en una sola variable, ignorando las otras, y esto puede reducir la calidad de las predicciones si las variables tienen correlación unas con otras.

La solución que utilizaremos es adaptar el modelo de regresión simple para que en lugar de tener una función de una sola variable, tengamos una función vectorial

$$\begin{aligned} Y = \beta_0 + \boldsymbol{\beta X} + \epsilon = \beta_0 + \begin{pmatrix}\beta_1 + \beta_2 + \ldots + \beta_n\end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} + \epsilon \\ = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon \end{aligned} \tag{8}$$

Donde X_i representa la iésima variable, y β_i cuantifica la asociación entre esa variable y la variable respuesta. La interpretación de β_i es el efecto medio sobre Y de un incremento de una unidad sobre la variable X_i, manteniendo fijos los demás parámetros.

Igual que ocurría con la regresión lineal simple, los coeficientes de regresión son desconocidos, y tenemos que estimarlos. Procedemos de la misma manera, minimizando el error cuadrático:

$$J(\hat{\beta}_0, \hat{\beta}_1, , \ldots, \hat{\beta}_n) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 X_1 - \hat{\beta}_2 X_2 - \ldots - \hat{\beta}_n X_n)^2$$

Al contrario que con la estimación simple, las fórmulas que nos permiten minimizar esta ecuación y obtener las estimaciones de los $n + 1$ parámetros son bastante complicadas, y, dado que las podemos obtener mediante cualquier software de estadística, no las incluiremos aquí.

Puede ser que de entre las variables independientes que estudiamos, haya algunas que no tengan relación con la variable independiente, incluso puede que sean todas. Para esto hacemos el contraste de hipótesis

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$$

$$H_1 : \exists i : \beta_i \neq 0$$

Para testear esta hipótesis utilizamos el estadístico F, que se define:

$$F = \frac{(n\sigma^2(Y) - J(\hat{\beta}_0, \hat{\beta}_1, , \ldots, \hat{\beta}_n))/p}{J(\hat{\beta}_0, \hat{\beta}_1, , \ldots, \hat{\beta}_n)/(n - p - 1)}$$

Si la suposición de linealidad es correcta y H_0 es cierta, se cumple

$$E[J(\hat{\beta}_0, \hat{\beta}_1, , \ldots, \hat{\beta}_n)/(n - p - 1)] = E[(n\sigma^2(Y) - J(\hat{\beta}_0, \hat{\beta}_1, , \ldots, \hat{\beta}_n))/p] \quad (9)$$

Por lo tanto, si el estadístico F tiene un valor lejos de 1, podemos rechazar H_0, y si no, no. ¿Cómo de lejos? Depende de n y p. Cuanto más grande sea n, menos lejos tendrá que estar, y al contrario con p. Cuando H_0 es cierta y los errores ϵ_i se distribuyen como una normal, el estadístico F sigue una distribución de Fisher-Snedecor, y podemos calcular el p-valor y utilizarlo para decidir si rechazamos H_0 o no. Cuanto más cerca de 0 esté el p-valor, mayor evidencia tenemos para rechazar H_0.

Hay veces que queremos comprobar si un subconjunto de q variables independientes no tienen relación con la variable dependiente. Para ello hacemos este contraste de hipótesis:

$$H_0 : \beta_{p-q+1} = \beta_{p-q+2} = \ldots = \beta_p = 0 \quad (10)$$

H_0 está escrita suponiendo que las q variables que queremos estudiar están al final del conjunto de todas las variables (y por tanto los coeficientes también).

Ahora, para un segundo modelo que utiliza todas las variables excepto las q últimas, tenemos:

$$F = \frac{(J(\hat{\beta}_0, \hat{\beta}_1, , \ldots, \hat{\beta}_{p-1}) - J(\hat{\beta}_0, \hat{\beta}_1, , \ldots, \hat{\beta}_n))/q}{J(\hat{\beta}_0, \hat{\beta}_1, , \ldots, \hat{\beta}_n)/(n-p-1)}$$

Este estadístico F nos dice el efecto de añadir estas q variables a un modelo en el que ya estaban todas las demás.

Sin embargo, parece que la información que nos da el estadístico F, ya la hemos obtenido en otro momento, al calcular cada coeficiente β_i como el incremento medio de la respuesta al incrementar en una unidad la variable X_i, estando fijas las demás, así que, ¿que nos aporta el estadístico F que no tengamos ya?

La respuesta la tenemos al observar un modelo con muchas variables: digamos $P = 100$. Vamos a suponer además, que ninguna de estas variables tiene relación alguna con la variable respuesta, es decir, $\beta_1 = \beta_2 = \ldots = \beta_p = 0$. En este caso, y teniendo en cuenta que el nivel de confianza es 0.05, hay una probabilidad del 5% de que un p-valor esté por debajo de 0.05, así que, de 100, podemos esperar encontrar 5 p-valores por debajo del 0.05, que es bastante cercano a 1, y que nos indicaría que los datos aportan una fuerte evidencia para rechazar H_0. Si es probable encontrar 5, podemos decir casi con total seguridad, que al menos habrá uno, y por lo tanto siempre fallaremos.

Esto no ocurre con el estadístico F, que tiene en cuenta la cantidad de variables, y haya la cantidad que haya, la probabilidad de obtener un p-valor inferior a 0.05 es del 5%. Tenemos un 95% de probabilidad de acertar.

Por último, y muy importante, el estadístico F funciona cuando p es "relativamente" pequeño (con 100 aún funciona bien), y mucho más pequeño si lo comparamos con n. Cuando p se acerca a n o incluso lo supera, tenemos que utilizar otras técnicas, y no podemos aplicar casi nada de lo que se ha visto hasta ahora.