Apuntes Tema 1

Albert Ribes Marzá

16 de octubre de 2017

Resumen

Los apuntes que vaya tomando en clase

1.

1.1. Introducción a ML

Ejemplo 1 Se pretende medir la temperatura (t) en un punto de una central nuclear, pero la temperatura es tan alta que no se puede medir directamente con ningún sensor. Se intentará deducir la temperatura así:

- t - temperatura a predecir (variable)
- x - vector de variables medibles que posiblemente inciden en t
- z - vector de variables **NO medibles** que posiblemente inciden en t

La relación completa es $t = \delta(x, z)$, que es una función.

Pero no conocemos z ➔Aun conociendo x, el valor de t oscila. La relación entre t y x se hace *estocástica*

$p(x, t)$ será la probabilidad de que con esa x se tenga esa temperatura.

Hay que construir una función $t = y(x)$ donde t sea el valor más plausible. El problema es que no conocemos p

La forma de atacar el problema será recolectar datos $\{(x_1, t_1), (x_2, t_2), \ldots, (x_n, t_n)\} = D$

$(x_n, t_n) \sim^{i.i.p.} p$ (variables independientes e identicamente distribuidas)

El objetivo del *Machine Learning* (ML) es obtener y a partir de D. En este caso es un problema de *regresión*

Ejemplo 2 Se tiene una planta de reciclaje, y se quieren clasificar los objetos que van pasando por la cinta. Los datos son:

- t - tipo de producto
- x - los atributos de los productos que captamos con una cámara
- z - los atributos que no captamos con la cámara

Es el mismo problema de antes, pero ahora t es discreta. Se trata entonces de un problema de *clasificación*.

1.2. Ejemplo introductorio: el ajuste polinómico

Tenemos $x \in \mathbb{R}$ y queremos predecir $t \in \mathbb{R}$ (*Regresión*)
En todos los problemas nos encontraremos con:

- $x_n \in (0,1)$, $x_n \sim U(0,1)$, (un conjunto de observaciones)
- $t_n = sin(2\pi x_n) + \varepsilon, \varepsilon \in N(0,\sigma^2)$, normalmente $\sigma^2 = 0{,}3^2$, y donde ε es el ruido aparentemente aleatorio, que proviene de los datos que no conocemos o de errores en la medición.

Vamos a intentar ajustar los datos. Sabemos que si los datos son continuos (no dan saltos) podemos ajustarlos con un polinomio en un intervalo:

- $P_n := \{c_0 + c_1 x + c_2 x^2 + \cdots + c_n x^n\} = \{\sum_{i=0}^{n} c_i x^i | c_i \in \mathbb{R}\}$
- $C \in \mathbb{R}^{n+1}$ son todos los parámetros.
- Llamamos $y(x;c) = \sum_{i=0}^{M} c_i x^i$ un modelo
- Respecto a X, y es una función no lineal
- Respecto a C, y es una función lineal

Diremos que un modelo es lineal cuando lo es respecto a los parámetros.
Ajustamos y a los datos $\{(x_1,t_1),(x_2,t_2),\ldots(x_n,t_n)\}$ definiendo una función de error (la función de error cuadrático):

$$E(c) = \frac{1}{2}\sum_{n=1}^{N}(y(x_n;C) - t_n)^2$$

Como E depende automáticamente de C, derivamos e igualamos a 0 para encontrar el mínimo. Hay que hacer n derivadas, una para cada c_j:

$$\frac{\partial E}{\partial c_j} = \frac{1}{2}\sum_{n=1}^{N} 2(y(x_n;C) - t_n)(x_n)^j$$

$$\frac{\partial E}{\partial c_j} = \sum_{n=0}^{N}\Big(\sum_{i=0}^{M}(c_i x_n^i - t_n)\Big)x_n^j = 0$$

El problema de todo esto es que no sabemos qué grado de polinomio deberíamos usar para reflejar el comportamiento de la variable.

- Si es demasiado pequeño no seremos capaces de ajustar la parte regular (y) de los datos (infra-ajuste)
- Si es demasiado grande se ajustará la parte regular (y) y también el ruido (sobre-ajuste)

Cómo elegir el grado del polinomio?

Únicamente conociendo $E(C)$ no se puede saber. Para hacerlo se usa una muestra alternativa de datos de validación. Esta muestra debería tener más datos.

Si observamos el error producido en nuestros datos con diferentes grados de polinomios, obviamente será más pequeño cuanto más grande sea el polinomio, puesto que tiene más flexibilidad. Pero si miramos qué ocurre con los datos de validación, veremos que al principio el error desciende, pero llega un punto en que empieza a subir. El punto mínimo se corresponde con el grado correcto.

El error empezará a subir porque el sobre-ajuste se ha adaptado a los datos aleatorios, pero en la muestra de validación no tienen por qué ser los mismos, y produce más error.

Si la muestra de validación tiene pocos datos, el mínimo estará poco definido, será más redondeado y más difícil de localizar.

Alternativa

Pero no siempre es posible tener suficientes datos de validación. Para esto hay una alternativa.

Uno se pregunta: ¿Si un polinomio de grado 9 "contiene" a los de grado más pequeño, no podría ocurrir que eligiéramos uno de grado mayor, y que él mismo anulara los coeficientes sobrantes hasta que sea del grado adecuado?

La respuesta es que espontaneamente esto no pasa, puesto que para igualar los datos aleatorios son necesarios coeficientes muy grandes. Si queremos que ocurra tenemos que forzarlo de alguna manera. Para hacerlo, redefiniremos nuestra función de error, de manera que también penalice los coeficientes demasiado grandes. Penalizaremos la norma 2, que equivale a la "distancia" pitagórica.

¿Pero cuanto tenemos que penalizarlo? Si nos pasamos o nos quedamos cortos no servirá de nada. Como no sabemos cuanto tenemos que penalizar, usaremos un parámetro λ que regulará la penalización que hacemos.

La función de error queda así:

$$E(C) = \frac{1}{2}\sum_{n=1}^{N}\Big(y(x_n;C) - t_n\Big)^2 + \frac{\lambda}{2}||c||^2$$

El $\frac{\lambda}{2}$ es simplemente para que al derivar quede más simple. Podría ser solo λ

1.3. Conceptos de inferencia estadística

$D = \{x_1, \dots x_n\}$ es una realización de una variable aleatoria (v.a.) X_n que tiene una función de distribución conocida $p(x_n;\theta), \theta \in \Theta$

Pero esa función de distribución tiene unos parámetros, y nos gustaría saber cuáles usar.

Por ejemplo, si se trata de una distribución normal, sabemos que nuestros datos se corresponden con $N(x, \mu, \sigma^2)$, pero viendo los datos no sabemos qué valores de μ y σ^2 deberíamos cojer para que se adaptaran lo más posible a los datos que tenemos.

Vamos a cojer los datos que maximicen nuestra verosimilitud (likelihood), esto es, los que hagan que sea más probable recoger los datos D

El objetivo es obtener una estimación $\hat{\theta}$ de θ, dado D

La probabilidad de recoger una muestra x_i es $p(x_i, \theta)$

Puesto que sabemos que los datos son i.i.d, la probabilidad de cojer todos los datos D es el producto de cada uno de ellos.

La probabilidad de obtener D es:

$$P_n(D, \theta) = \prod_{n=1}^{N} p(x_n, \theta)$$

Definimos la función de verosimilitud (likelihood) así:

$$\mathcal{L} : \theta \to \mathbb{R}$$

$$\theta \to \mathcal{L}(\theta) = P_n(D; \theta)$$

Es decir, es una función que indica cómo de probable es haber recogido los datos D usando los parámetros θ.

Elegiremos los paráetros θ que maximicen esa probabilidad, los que maximicen $P(D, \theta) = \prod_{n=1}^{N} p(x_n, \theta)$

El estimador de máxima verosimilitud es $\hat{\theta} = \text{argmax } \mathcal{L}(\theta)$, $\theta \in \Theta$

Si es de una sola variable, la forma de hacerlo es derivar e igualar a 0.

Es conveniente operar con el logaritmo de α, pues simplifica la maximización de un producto:

$$\ln(p_1 p_2 \dots p_n) = \ln(p_1) + \ln(p_2) + \dots + \ln(p_n)$$

Ejemplo

Tenemos un conjunto de datos $D = \{x_1, x_2, \dots, x_n\}$ que siguen una distribución normal $X_n \sim N(x_n, \mu, \sigma^2)$

Se sabe que la función de densidad de la distrubión normal es:

$$N(x, \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Entonces la probabilidad de haber cogido los datos D es:

$$P(D, \mu, \sigma^2) = \prod_{i=1}^{n} \Big(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \Big)$$

Pero para hacer los cálculos más sencillos trabajaremos con el logaritmo, que no cambia los máximos relativos. Además, le vamos a cambiar el signo, y ya no buscaremos maximizarlo, sino minimizarlo.

$$l = -\ln(P(D, \mu, \sigma^2)) = -\ln \prod_{i=1}^{n} \Big(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \Big) = -\sum_{i=1}^{N} \ln \big(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \big)$$

Se simplifica así:

$$l = -\sum_{i=1}^{N} \ln \big(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \big)$$

$$l = -\sum_{i=1}^{N} \Big[\ln\big(\frac{1}{\sigma\sqrt{2\pi}}\big) + \ln\big(e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\big) \Big]$$

$$l = -\sum_{i=1}^{N} \Big[-\ln\big(\sigma\sqrt{2\pi}\big) - \frac{(x_i-\mu)^2}{2\sigma^2} \Big]$$

$$l = \sum_{i=1}^{N} \Big[\ln\big(\sigma\sqrt{2\pi}\big) + \frac{(x_i-\mu)^2}{2\sigma^2} \Big]$$

Ahora derivamos respecto de μ y de σ^2 e igualamos a 0 para encontrar los extremos:

$$\frac{\partial l}{\partial \mu} = \sum_{i=1}^{N} \Big[0 + \frac{1}{2\sigma^2} \cdot 2(x_i-\mu)(-1) \Big]$$

$$\frac{\partial l}{\partial \mu} = \sum_{i=1}^{N} \Big[-\frac{x_i-\mu}{\sigma^2} \Big]$$

$$\frac{\partial l}{\partial \mu} = -\frac{1}{\sigma^2} \sum_{i=1}^{N} \Big[x_i - \mu \Big] = 0$$

$$\sum_{i=1}^{N} x_i - \sum_{i=1}^{N} \mu = 0$$

$$\sum_{i=1}^{N} x_i - N\mu = 0$$

$$\sum_{i=1}^{N} x_i = N\mu$$

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Y respecto de σ^2:

$$l = \sum_{i=1}^{N} \Big[\ln\big(\sigma\sqrt{2\pi}\big) + \frac{(x_i-\mu)^2}{2\sigma^2} \Big]$$

$$\frac{\partial l}{\partial \sigma^2} = \sum_{i=1}^{N} \Big[\frac{1}{\sigma\sqrt{2\pi}} \cdot \frac{\sqrt{2\pi}}{2\sigma} + \frac{(x_i-\mu)^2}{2}\big(-\frac{1}{\sigma^4}\big) \Big]$$

$$\frac{\partial l}{\partial \sigma^2} = \sum_{i=1}^{N} \Big[\frac{1}{2\sigma^2} - \frac{(x_i-\mu)^2}{2\sigma^4} \Big] = 0$$

$$\sum_{i=1}^{N} \Big[\frac{1}{2\sigma^2} \Big] - \sum_{i=1}^{N} \Big[\frac{(x_i-\mu)^2}{2\sigma^4} \Big] = 0$$

$$\frac{N}{2\sigma^2} = \frac{1}{2\sigma^4}\sum_{i=1}^{N}(x_i - \mu)^2$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

De este modo hemos encontrado las ecuaciones clásicas para encontrar la media y la varianza muestral, que son la media y varianza más probables teniendo en cuenta los datos que hemos recogido.

Se podría aplicar el mismo procedimiento con alguna otra distribución.

Nota: habría que asegurarse que realmente se trata de mínimos mirando la segunda derivada y comprobando que da valores positivos. Se deja como ejercicio

1.4. Propiedades de un estimador

2. Reducción de dimensión

Para hacer *Machine Learning* es interesante tener los datos lo más simplificados posible, pues eso evita el sobre-ajuste. Existen muchos métodos para reducir la dimensión de un problema. Reducir la dimensión se podría entender como quedarse con una sombra de la imagen real que tenemos. Esto es: si todos los datos que tenemos estuvieran en 3 dimensiones, podría interesarnos trabajar con la sombra que proyectan esos datos, de manera que trabajaríamos con solo 2 dimensiones.

Pues hacemos los mismo, pero con muchas dimensiones.

Las ventajas que tiene esto son que evita el sobre-ajuste, nos permite entender mejor los datos y que son más fáciles de representar, con plots o dibujos.

Pero hay que cojer una buena proyección de los datos reales. Puesto que está claro que vamos perder información (datos, en realidad), cojeremos una proyección que refleje lo que nos interesa, y que deseche otras cosas.

Es por eso que hay muchas formas de reducir la dimensión de un conjunto de datos, cada una según la prioridad que uno tenga, y cogiendo las proyecciones más adecuadas para cada necesidad.

Ahora veremos algunas de las formas de reducir la dimensión:

2.1. Principal Components Analisis (PCA)

Este algoritmo tiene como prioridad preservar la varianza de los datos, maximizar la dispersión en las proyecciones.

Esto es, en la analogía de la sombra, cojer la sombra que tenga más area.

De forma más técnica:

Tenemos una muestra de datos $\{X_1, X_2, \ldots, X_n\}, X_i \in \mathbb{R}^d$ que provienen de un vector aleatorio $X = \{X_1, \ldots, X_n\}^T$. Cada una de las X_i es una variable (aleatoria?) y tenemos d muestras en cada una de las variables.

Disponemos también de la matriz de covarianzas Σ.

La matriz de covarianzas es una matriz de $n \times n$ donde Σ_{ij} es $\text{var}(X_i, X_j)$ si $i \neq j$ y Σ_{ii} es σ_i^2

Tenemos datos en n dimensiones, y decidimos que queremos únicamente k dimensiones, $k < n$, y no cualesquiera dimensiones, sino las que maximicen la varianza.

Hemos de encontrar entonces k vectores n-dimensionales. Encontraremos n vectores que serán todos "perpendiculares" entre ellos y cojeremos los k vectores que tengan más varianza.

Nuestro objetivo entonces obtener un nuevo sistema de coordenadas $Y = (Y_1, \ldots, Y_n)$ que cumpla estas condiciones:

1. $\text{Covar}(Y_i, Y_j) = 0$ si $i \neq j$
2. $\text{Var}(Y_1) > \text{Var}(Y_2) > \ldots > \text{Var}(Y_n)$ (de hecho los ordenaremos decrecientemente)
3. $\sum_{i=1}^{d} Var(X_i) = \sum_{i=1}^{d} Var(Y_i)$

Encontraremos la proyección Y_i encontrando un vector w_i que cumpla que $Y_i = w_i^T \cdot X$

Como hay muchos vectores que cumplen esa condición (vectores que tienen todos la misma dirección, pero distinto módulo) establecemos la condición sobre w_i de que la norma 2 al cuadrado sea 1, esto es: $||w_i||_2^2 = 1 \Rightarrow w_{i1}^2 + w_{i2}^2 + \cdots + w_{in}^2 = 1$

Objetivo: w_1 ha de maximizar la varianza de Y_i, sujeto a que $||w_i|| = 1$

$$Var(Y_i) = Var(w_i^T \cdot X) = w_1^T \cdot Var(X) \cdot w_i = w_1^T \cdot \Sigma \cdot w_i$$

Este último paso es algo que se sabe y que sale en Wikipedia. Nos lo creemos.

Para resolver un problema de maximización sujeto a algunas condiciones se hace con el método de los multiplicadores de Lagrange.

Anexo: método de Lagrange

El método de Lagrange sirve para encontrar los extremos (máximos o mínimos) que hay en una función sujeto a algunas condiciones de igualdad. La función puede tener una cantidad arbitraria de parámetros, y se acepta también una cantidad arbitraria de condiciones.

Si tenemos la función $f(x_1, \ldots, x_n)$ y las condiciones $g_1(x_1, \ldots, x_n) = 0, \ldots, g_m(x_1, \ldots, x_n) = 0$, definimos nuevas variables λ (habrá una por cada condición que haya) y construimas la función de Lagrange:

$$\mathcal{L}(x_1, \ldots, x_n, \lambda_1, \ldots, \lambda_m) = f(x_1, \ldots, x_n) - \sum_{k=1}^{m} \lambda_k g_k((x_1, \ldots, x_n))$$

En el caso particular de una función de dos parametros $f(x, y)$ y una restricción $g(x, y) = 0$ sería así:

$$\mathcal{L}(x, y, \lambda) = f(x, y) - \lambda g(x, y)$$

El siguiente paso es derivar $\mathcal{L}$ sobre cada uno de los parámetros (tanto los reales como los añadidos) e igualar todas las ecuaciones a 0. Eso dará como resultado un sistema de ecuaciones que tendrá tantos resultados como extremos existan cumpliendo todas las condiciones.

Ahora ya únicamente queda mirar todos esos puntos y decidir cuales son máximos o mínimos.

Fin del Anexo: método de Lagrange

Entonces, nuestro problema es maximizar

$$w_i^T \cdot \Sigma \cdot w_i$$

sujeto a que

$$\sum_{j=1}^{d} \left(w_{ik}^2\right) - 1 = 0$$

Construimos la función de Lagrange:

$$\mathcal{L}(w_i, \lambda) = w_i^T \cdot \Sigma \cdot w_i - \lambda\Big(\sum_{j=1}^{d} \left(w_{ik}^2\right) - 1\Big)$$

Y derivamos e igualamos a 0:

(No tengo muy claro cómo se hace esa derivada, pero nos creemos que es así)

$$\frac{\partial \mathcal{L}}{\partial w_i} = 2\Sigma \cdot w_i - 2\lambda w_i = 0$$

$$\Sigma \cdot w_i = \lambda w_i$$

La expresión que nos ha quedado se corresponde con la definición de vector y valor propio de una matriz (*eigenvector* y *eigenvalue*). Se dice que λ es un valor propio de la matriz Σ si existe un vector w tal que:

$$\Sigma \cdot w = \lambda w$$

Al vector w se le llama vector propio de Σ

Vemos entonces que los vectores de proyección w_i que estamos buscando se corresponden con los vectores propios de la matriz Σ.

Pero hemos exigido que los vectores de proyección esten ordenados decrecientemente por varianza, i. e. el primer vector de proyección w_1 debe ser el que tenga más varianza en la proyección, el segundo w_2, etc.

Veamos cuál será la varianza del vector w_i

$$Var(Y_i) = Var(w_i^T \cdot \Sigma) = w_i^T \cdot \Sigma \cdot w_i$$

Pero hemos visto que

$$\Sigma \cdot w_i = \lambda w_i$$

Por lo tanto podemos sustituir:

$$w_i^T \cdot \Sigma \cdot w_i = w_i^T \lambda w_i = \lambda w_i^T w_i$$

Y como hemos exigido que $w_i^T w_i = 1$, tenemos que la varianza será λ

Entonces, el orden en que hemos de cojer los vectores propios es respecto a su valor propio: primero el vector con mayor valor propio, etc.

Pero en realidad también hemos exigido que los vectores de proyección elegidos también sean perpendiculares entre ellos. Para conseguir esto, haremos Lagrange para encontrar únicamente el primero de los vectores de proyección, y luego, para encontrar el segundo hacemos igual pero estableciendo también la condición de que el nuevo vector de proyección sea perpendicular con el primero, i, e: $w_1^T w_2 = 0$

Cuando tenemos todos los vectores de proyección $w_1, w_2, \ldots, w_n$, podemos definir la matriz

$$A = \begin{bmatrix} w_{11} & w_{21} & \ldots & w_{d1} \\ w_{12} & w_{22} & \ldots & w_{d2} \\ \vdots & \vdots & \ddots & \vdots \\ w_{1n} & w_{2n} & \ldots & w_{dn} \end{bmatrix}$$

Y entonces podemos decir que nuestros nuevos datos son:

$$Y = A^T X$$

Ahora es cuando hemos de decidir con cuántos de los componentes principales nos quedamos. Si nos los quedamos todos lo único que habremos hecho será una transformación de los datos.

Si queremos quedarnos con m componentes principales y en total había d ($m \leq d$), entonces

$$\frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{d} \lambda_i} \times 100$$

Es el porcentaje de la varianza con el que nos estamos quedando

2.2. Fisher Discriminant Analysis (FDA)

Queda pendiente para hacer esta sección.

Tiene que ver con hacer otro modelo para reducir la dimensión, dando prioridad a otras cosas

3. Clustering

Clustering quiere decir agrupar. Un algoritmo de clustering recibe como entrada un conjunto de datos, y él clasifica esos datos en varios grupos, en varios clusters, en función de algún parámetro.

Normalmente la cantidad de clusters es arbitraria, decidida por factores ajenos al problema. De hecho es un problema complicado decidir cuál es la cantidad adecuada de clusters, y existen algunos algoritmos que intentan calcularla, pero son poco genéricos.

Para los algoritmos que trataremos aquí supondremos que la cantidad de clusters ya viene dada. Se tratará de encontrar "dónde están" esos clusters y ver qué elementos pertenecen a cada uno de los clusters. Como veremos, la mayoría de problemas problemas de optimización de clustering son NP-completos, y por eso buscaremos una buena solución en vez de la mejor solución.

3.1. Algoritmo de K-means

Disponemos de un conjunto de datos $D = \{x_1, x_2, \ldots, x_n\}, x_i \in \mathbb{R}^d$ y queremos encontrar K prototipos (centroides) $P = \{\mu_1, \mu_2, \ldots, \mu_K\}, \mu_i \in \mathbb{R}^d$ a los que asociar los datos D

El criterio será minimizar la suma de distancias de cada dato x_i a su cluster más cercano, que será el cluster al que esté asignado.

Definimos una variable binaria π_{ik}:

$$\pi_{ik} = \begin{cases} 1 & \text{si el dato } x_i \text{ está asignado al cluster } k \\ 0 & \text{otherwise} \end{cases}$$

Formalmente el criterio a minimizar será:

$$J = \sum_{i=1}^{N} \sum_{k=1}^{K} \pi_{ik} ||x_i - \mu_k||^2$$

Está demostrado que dado un conjunto de datos, encontrar la "posición" de K clústers y la asignación de datos a clusters de la manera más optima es un problema NP-completo, de modo que para resolver esto vamos a tener un problema.

Sin embargo, resolver la mitad del problema es fácil: si te dicen a priori la "posición" de cada uno de los clusters, es fácil ver qué datos hay que asignar a cada cluster. Si por el contrario te dicen qué datos irán asignados al mismo cluster, es fácil encontrar la posición de cada uno de los clusters.

Entonces para resolver este problema, partiremos de una de las mitades del problema, que se habrá encontrado arbitrariamente, y entonces se buscará la otra mitad más óptima. A continuación se recalculará la primera mitad en función de estos datos, y este proceso se repetirá hasta que ya no se mejore más.

De esta manera terminamos teniendo un máximo local, pero no absoluto. La calidad de la solución dependerá de los datos iniciales. Se han hecho muchos estudios sobre qué datos iniciales son más adecuados, pero al final parece que lo mejor es usar datos aleatorios de entre el conjunto de datos D.

Este algoritmo es extremadamente rápido, por lo que normalmente se ejecuta varias veces con datos iniciales distintos y se mantiene el que da mejores resultados.

Ahora faltaría ver cómo se encuentra la segunda mitad. Queda para más adelante.

3.2. Mezcla de Gaussianas (Mixture of Gaussians MoG)