Conversaciones en línea sobre Ciencia de Datos

Escuela de Matemáticas Bourbaki

Marzo 2020

1 Introducción

Este curso en línea está organizado en forma de charlas. El objetivo es introducir y utilizar el lenguaje de las matemáticas para estudiar algunas de las ideas y dificultades recurrentes en la Ciencia de Datos.

La intención del curso es ofrecer una perspectiva general sobre diversos aspectos en la Ciencias de Datos. El curso esta dividido en dos partes fundamentales, cada una consta de tres clases:

- Eficiencia computacional y estadística
- Teoría de la información e inteligencia artificial

1.1 Temario

Los temas a tratar son los siguientes:

- ¿La bendición o la maldición de la dimensión en la clasificación lineal?
- Detección de anomalías mediante el análisis topológico de datos.
- Consideraciones matemáticas de la computación cuántica.
- Codificación de la información y Entropía
- Inteligencia artificial v.s. los grandes maestros de los juegos de estrategia
- Programación dinámica, una invitación al Aprendizaje por refuerzo

1.2 Estructura del curso

Cada una de las clases está dividida en dos partes:

- Los primeros 45 minutos enseñaremos los detalles matemáticos de algún modesto aspecto relacionado con el tema a tratar, buscamos que los asistentes se expongan a una explicación minuciosa y esta les sea útil para entender mejor la segunda parte de la clase. Compartiremos con los asistentes unas notas cortas sobre esta sección.
- Los segundos 45 minutos se dedicarán a tratar aspectos más generales y elaborados del tema, incluso algunos más polémicas.

2 La maldición v.s. la bendición de la dimensión

Uno de los aspectos más importantes en Machine Learning es la batalla entre la cantidad de ejemplos de una base de datos y la dimensión donde vive alguna vectorización de estos ejemplos.

Las llamadas bendición y maldición de la dimensión se manifiestan cuando la cantidad de ejemplos es menor a la cantidad de dimensiones del problema vectorizado.

2.1 Primera parte

2.1.1 Problemas supervisados de clasificación binaria

Supongamos que estamos en un problema clásico de aprendizaje supervisado tipo clasificación binaria en $\mathbb{R}^d$ donde buscamos entrenar a nuestro modelo con un conjunto de ejemplos $S = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ i.e. $x_i \in \mathbb{R}^d, y_i \in \{-1, +1\}$.

El objetivo de un algoritmo que se entrene utilizando a S es encontrar una función $f^* : \mathbb{R}^d \to \{-1, 1\}$ que cumpla lo siguiente:

1. (Entrenamiento) Para la mayor parte de los puntos $(x_i, y_i) \in S$ la función f^* satisfaga $f^*(x_i) = y_i$. Es importante notar que algunas veces no es posible que la igualdad anterior sea para todos los puntos en S sin embargo por lo menos nos gustaría que fuera cierto para buena parte de ellos.

2. (Predicción) Un posible supuesto en un problema de Machine Learning es que exista alguna regla que de manera general pueda aplicarse a una observación x y nos regrese un resultado y. Este supuesto se traduce en que exista una función desconocida $f : \mathbb{R}^d \to \mathbb{R}$ (que nuevamente para la gran mayoría de los $(x_i, y_i) \in S$ satisfaga $f(x_i) = y_i$). El segundo objetivo (y quizás el más importante en Machine Learning) es que f y f^* sean parecidas.

2.1.2 Una maldición de la dimensión en un problema de clasificación

Para simplificar el estudio y poder hablar de árboles de decisión de la manera más sencilla posible vamos a simplificar aún más un problema de clasificación y por el momento supondremos que $x_i \in \{0, 1\}^d$. Esto significa que las características de nuestra base de datos son binarias, pensemos por ejemplo en una imagen que solo tiene pixeles negros o blancos. Eso significa que utilizando nuestro conjunto S buscamos aproximar alguna función $f : \{0, 1\}^d \to \{-1, +1\}$.

Definition 2.1. Un árbol de decisión de profundidad r y dimensión d es un conjunto de r-preguntas ordenadas sobre un vector $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ de la siguiente forma:

- x_{i_1} es igual a cero o a uno?
- x_{i_2} es igual a cero o a uno?
- x_{i_3} es igual a cero o a uno?

...

- x_{i_r} es igual a cero o a uno?

De tal forma que cada hoja formada por posibles respuestas a todas estas preguntas nos arrojan una posible clasificación ya sea -1 o $+1$.

Exercise 2.1. *Sea $S = \{(0, 0, 1), (0, 1, -1), (1, 0, -1), (1, 1, 1)\}$. Encontrar el árbol de clasificación que prediga la clasificación en S. Es posible encontrar un árbol de clasificación de profundidad uno que también prediga la clasificación en S?*

La ventaja y al mismo tiempo gran problema de los árboles de decisión es la siguiente:

Proposition 2.2. *Toda función $f : \{0, 1\}^d \to \{-1, +1\}$ corresponde con un árbol de decisión con profundidad d.*

Lo anterior genera un gran problema debido a la posibilidad de incurrir en sobre-ajuste, más adelante en la clase número tres de este curso hablaremos sobre la navaja de Occman que es un resultado que previene el sobre-ajuste, el problema es que estamos considerando árboles de decisión con profundidad demasiado grande. Para explicar con formalidad las desventajas causadas por ese sobre-ajuste podemos enunciar el siguiente resultado.

Theorem 2.3. *La cantidad de ejemplos necesaria para entrenar un árbol de decisión de manera eficiente crece de manera proporcional a* 2^d.

El resultado anterior se puede entender como una maldición de la dimensión pues a medida que d crezca, la cantidad de ejemplos necesarios 2^d podría ser excesivamente grande.

Por el momento no hemos hablado del algoritmo que entrena de manera eficiente un árbol de decisión, solo mencionamos que está relacionado con la idea de simpleza proveniente de Occman. En el curso 4 hablaremos de la teoría de información y su importancia para entrenar un árbol de decisión.

2.1.3 Una bendición de la dimensión en un problema de clasificación

Una de las formas más simples para afrontar un problema de clasificación es mediante la llamada clasificación lineal. Para esto podemos regresar sin problema al caso cuando $x_i \in \mathbb{R}^d$.

Definition 2.2. El producto punto entre dos vectores $x = (x_1, \ldots, x_d), x' = (x'_1, \ldots, x'_d) \in \mathbb{R}^d$ se define como $x \cdot x' = x_1 x_1 + \ldots + x_d x'_d$

Exercise 2.4. *Verificar que el producto punto de dos vectores coincide con multiplicar dos matrices, cuáles?*

Definition 2.3. Dos vectores $x, x' \in \mathbb{R}^d$ son linealmente separables mediante el hiperplano definido por $\beta = (\beta_1, \ldots, \beta_d) \in \mathbb{R}^d$ y $\beta_0 \in \mathbb{R}$ si $x \cdot \beta + \beta_0 > 0$ y $x \cdot \beta + \beta_0 < 0$. Decimos que dos conjuntos S, S' de vectores en $\mathbb{R}^d$ son linealmente separables mediante β, β_0 si todo par de puntos en S y S' son linealmente separables.

Exercise 2.5. *Partiendo de la ecuación clásica de una recta* $y = mx + b$ *escribir qué significa que dos puntos en el plano* $\mathbb{R}^2$ *sean separables linealmente, preferentemente utilizando la noción de producto punto.*

En esta sección propondremos una condición algebráica sobre el conjunto S para garantizar que es separable linealmente por un hiperplano sin importar cómo sus elementos estén clasificados. Esto es considerablemente más fuerte a lo que buscábamos en un inicio.

La condición algebraica necesaria se define a continuación:

Definition 2.4.

- Una familia de vectores $x_1, x_2, \ldots x_m \in \mathbb{R}^d$ son linealmente dependientes si existen $\gamma_1, \gamma_2, \ldots, \gamma_m \in \mathbb{R}$ distintos de cero tales que $\gamma_1 x_1 + \ldots \gamma_m x_m = 0$
- Sea $S \subseteq \mathbb{R}^d$ un conjunto arbitrario de puntos de tamaño N (notemos que esta vez no estamos suponiendo que los puntos están etiquetados). Decimos que S está en posición general cuando no existe un subconjunto $S' \subset S$ de tamaño menor a d que sea linealmente dependiente.

Exercise 2.6. *Cuál es la mayor cantidad de puntos en* $\mathbb{R}^2$ *que pueden estar en posición general? Y en* $\mathbb{R}^3$ *o* $\mathbb{R}$*?*

Theorem 2.7. *(Cover) sea* $S \subseteq \mathbb{R}^d$ *un conjunto arbitrario de tamaño* N *con* $d > N$. *Además supondremos que* S *está en posición general en el sub-espacio generado por ellos. Entonces cualquier etiquetado de* S: $\left(x_{i_j}, -1\right)$ *si* $j \leq m \leq N$ *y* $\left(x_{i_j}, 1\right)$ *si* $j > m$ *es linealmente separable.*

El trabajo de Cover es mucho más profundo que el resultado anterior y vale la pena consultar sus detalles en el artículo originalGeometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition.

2.2 Segunda parte

En la segunda parte nos concentraremos únicamente en las llamadas bendiciones de la dimensión. La base técnica para estos resultados son los llamados Teoremas de Concentración en Probabilidad, comenzaremos esta sección describiendo algunos de estos resultados desde un aspecto púramente teórico.

2.2.1 Teoremas de Concentración

La mayoría de estos resultados son suficientemente antiguos sin embargo nos concentraremos en dos resultados particularmente, uno de Chebyshev y el otro de Lévy.

Lemma 2.8. *(Desigualdad de Chebyshev) Sea X una variable aleatoria con valores en $\mathbb{R}$ y $\mathbb{E}[X] = \mu, Var(X) = \sigma^2$, entonces para cualquier $t > 0$ tenemos:*

$$\mathbb{P}(|X - \mu| > t) < \frac{\sigma^2}{t^2}$$

Esta desigualdad es considerablemente más poderosa cuando la dimensión donde toman valores las variables aleatorias crece considerablemente, aún así en la siguiente sección daremos un ejemplo que nos da una idea de cómo este tipo de desigualdades son ejemplos de la llamada bendición de la dimensión.

Antes de continuar vamos a introducir una clase de funciones súmamente importantes, también en el contexto de Ciencias de Datos.

Definition 2.5. Sea $f : \mathbb{R}^n \to \mathbb{R}^m$ y $\nu > 0$, decimos que f es una función ν-Lipschitz cuando para $x, y \in \mathbb{R}^n$ se tiene:

$$||f(x) - f(y)||_2 \leq \nu ||x - y||_2$$

La intuición de estas funciones para un Científico de Datos es la siguiente:

Utilizando el acercamiento Bayesiano supongamos que tenemos una distribución de probabilidad D sobre $\mathbb{R}^d \times \{-1, +1\}$.

Sean X, Y variables aleatorias en $\mathbb{R}^d$ y $\{-1, +1\}$ cuyas distribuciones coinciden con las distribuciones marginales de D.

Definamos a la función f de la siguiente manera:

$$f(x) = \mathbb{P}_D(Y = 1|X)$$

El acercamiento Bayesiano consiste en aproximar de la mejor manera la función f utilizando una base de datos S.

Exercise 2.9. *¿Qué repercución tiene en Ciencia de Datos que la función f sea ν-Lipschitz?*

En el estudio de la maldición de la dimensión en el contexto de K-NN las funciones Lipschitz son muy importantes.

Exercise 2.10. *Fijemos un vector $\beta \in \mathbb{R}^d$, definimos $f : \mathbb{R}^d \to \mathbb{R}$ de la siguiente forma:*

$$f(x) = \beta \cdot x$$

Demostrar que f es 1-Lipschitz.

Definition 2.6. La esfera de dimensión $d - 1$ es el conjunto de puntos

$$\mathbb{S}_{d-1} = \{x \in \mathbb{R}^d : ||x||_2 = 1\}$$

Proposition 2.11. *(Lévy) Sea X una variable aleatoria con valores en la esfera $\mathbb{S}_{d-1}$, cuya distribución es simétrica respecto a rotaciones y $f : \mathbb{R}^d \to \mathbb{R}$ una función ν-Lipschitz. Entonces para cualquier $t > 0$*

$$\mathbb{P}(|f(X) - \mathbb{E}[f(X)]| > t) < 2 \cdot exp\left(-\frac{d \cdot t^2}{2c\nu^2}\right)$$

Es importante notar que esta desigualdad será más y más precisa a medida que d sea suficientemente grande dado el decrecimiento de la función exponencial para valores negativos.

2.2.2 Algunas aplicaciones de los teoremas de concentración

- Comencemos con una sencilla pero agradable aplicación de la desigualdad de Chebyshev al estudio probabilista del lanzamiento de una moneda:

  Supongamos que lanzamos una moneda con águila y sol d-veces. ¿Cuál es la probabilidad de obtener águila $\frac{3}{4}d$-veces?

  Si le llamamos A_d al número de veces que obtenemos águila después de lanzar la moneda d-veces entonces entonces es fácil convencerse de que:

  Exercise 2.12. *Definir el espacio de probabilidad adecuado y calcular* $\mathbb{E}[A_d] = \frac{d}{2}$, $Var(A_d) = \frac{d}{4}$.

  Al sustituir en la desigualdad de Chebyshev $t = \frac{d}{4}$ obtenemos:

  $$\mathbb{P}\left(|A_d - \frac{d}{2}| > \frac{d}{4}\right) < \frac{4}{d}$$

- Una segunda aplicación del fenómeno de concentración de medidas en altas dimensiones es el llamado lema de Johnson-Lindenstrauss el cuál puede pensarse como un lema para reducir la dimensión.

  Antes de eso recordemos un resultado importante sobre la reducción de la dimensión en Ciencia de Datos.

  Supongamos que tenemos un conjunto $S = \{x_1, \ldots, x_N\} \subseteq \mathbb{R}^d$. La reducción de la dimensión consiste en encontrar puntos $S' = \{x'_1, \ldots, x'_N\} \subseteq \mathbb{R}^p$ con p "considerablemenente" más pequeño que d de tal forma que las dos siguientes condiciones se cumplan:

  1. La geometría del conjunto S' se parezca lo más posible a la geometría del conjunto S.
  2. Exista una transformación lineal M que envíe S a S'.

  El primer punto es clave pues no hemos sido suficientemente claros sobre qué significa la similitud geométrica.

  En el contexto de PCA lo que significa esa similitud geométrica es lo siguiente:

  No solo buscamos una transformación lineal M sino además queremos una transformación M' que "regrese" a S, eso quiere decir que minimice la distancia promedio entre $S'' = M'(M(S))$ y S. No es difícil ver que este acercamiento es equivalente a buscar maximizar la varianza del conjunto S'. Normalmente el resultado cuando p es demasiado pequeño respecto a d es poco satisfactorio pues el conjunto S' será poco representativo del conjunto inicial S.

  Una manera alternativa preservar la geometría del conjunto S que por cierto podría ser relevante cuando se necesite utilizar un algoritmo de proximidad es el llamado lema de Johnson-Lindenstrauss el cuál dice lo siguiente:

  Lemma 2.13. *(Johnson-Lindenstrauss) Sea* $S = \{x_1, \ldots, x_N\} \subseteq \mathbb{R}^d$ *y* $\epsilon > 0$. *Si* $p \geq \frac{C}{\epsilon} log(N)$ *entonces con alta probabilidad existe un subespacio vectorial (aleatorio)* $\mathbb{R}^p \subseteq \mathbb{R}^d$ *y una transformación lineal(también aleatoria)* $M : \mathbb{R}^d \to \mathbb{R}^p$ *tal que para todo par de puntos* $x_i, x_j \in S$*:*

  $$(1-\epsilon)||x_i - x_j||_2 \leq ||M(x_i) - M(x_j)||_2 \leq (1+\epsilon)||x_i - x_j||_2$$

  La precisión en probabilista de este resultado depende de que d sea suficientemente grande utilizando el resultado de Lévy.

- Ahora recordemos el resultado de Cover que era nuestro primer ejemplo de la bendición de la dimensión, si $d > N$ entonces cualquier partición de cualquier conjunto en posición general de tamaño N es linealmente separable por un hiperplano que pasa por el orgien. Una pregunta inmediata es ¿qué pasa cuando la desigualdad anterior no se cumple? Es decir supongamos que $N > d$. Este es nuestro ejemplo final de una bendición de la dimensión: los teoremas estocásticos de separación lineal nos dicen que si d es suficientemente grande entonces con alta probabilidad (dependiendo de lo grande que sea d) un conjunto S generado por variables aleatorias identicamente distribuidas e independientes es linealmente separable. En este caso la noción probabilista de i.d.d. es el análogo de estar en posición general.

- Otra importante aplicación de los teoremas de concentración a la Ciencia de Datos es una modificación de la selección del modelo de Akaike, es decir un tipo de regularización.