UNIVERSIDAD
DE GRANADA

Aprendizaje Automático
grado en ingeniería informática

TRABAJO 1

Cuestiones de Teoría

Autor
Vladislav Nikolov Vasilev

Rama
Computación y Sistemas Inteligentes

Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación

Curso 2018-2019

Índice

Ejercicio 1

Identificar, para cada una de las siguientes tareas, cuál es el problema, qué tipo de aprendizaje es el adecuado (supervisado, no supervisado, por refuerzo) y los elementos de aprendizaje $(\mathcal{X}, f, \mathcal{Y})$ que deberíamos usar en cada caso. Si una tarea se ajusta a más de un tipo, explicar como y describir los elementos para cada tipo.

a) Clasificación automática de cartas por distrito postal.

Solución

El problema ante el que nos encontramos en este caso consiste clasificar cartas según su distrito postal. Posiblemente sea un problema de clasificación donde hayan k clases, una por cada código postal, así que el aprendizaje supervisado puede ser la opción más efectiva para determinar a qué clase pertenece una carta.

El conjunto de datos de entrada $\mathcal{X}$ puede ser por ejemplo los datos del destinatario (su dirección, por ejemplo). El conjunto de etiquetas $\mathcal{Y}$ que podríamos usar son k etiquetas de números enteros, como por ejemplo $\{0, 1, \ldots, k\}$, cada una de las cuáles se asocia con un distrito postal. La función f sería alguna función tal que $f : \mathcal{X} \mapsto \mathcal{Y}$.

b) Decidir si un determinado índice del mercado de valores subirá o bajará dentro de un período de tiempo fijado.

Solución

El problema en este caso consiste en predecir o decidir a partir de unos datos de entrada una clase (la de si subirá o bajará el índice de mercado). Por tanto este problema se puede ver como una clasificación binaria $(0, 1)$ o $(-1, 1)$.

En el caso de los datos de entrada $\mathcal{X}$ podríamos utilizar valores del mercado y el tiempo. En el caso de los datos de salida o etiquetas $\mathcal{Y}$ podríamos tener las etiquetas $(-1, 1)$, siendo -1 el caso de bajar el índice y 1 el de subir. Y por último, f sería una función que relacione a $\mathcal{X}$ y a $\mathcal{Y}$ tal que $f : \mathcal{X} \mapsto \mathcal{Y}$.

c) Hacer que un dron sea capaz de rodear un obstáculo.

Solución

El problema en este caso es hacer que un dron aprenda a esquivar un obstáculo rodeándolo. Como el objetivo no es clasificar ninguna información, ni predecir ningún valor real ni buscar características o patrones en los datos, parece que el tipo de aprendizaje más adecuado es el aprendizaje por refuerzo. Esto se podría hacer mediante un simulador en un ordenador, donde se representaría el espacio donde se quiere entrenar al dron. Una vez entrenado en este simulador, se podría transferir todo lo aprendido al dron y ver cómo se desempeña.

Como tal, el aprendizaje por refuerzo no tendría ni entradas $\mathcal{X}$, ni etiquetas de salida $\mathcal{Y}$ ni una función f para el aprendizaje, pero sí que tendría otra información que se correspondería con un Proceso de Decisión de Markov (MDP), como por ejemplo un conjunto de estados, acciones, probabilidades de transicionar de un estado a otro, una recompensa por transicionar de estado, etc.

d) Dada una colección de fotos de perros, posiblemente de distintas razas, establecer cuántas razas distintas hay representadas en la colección.

Solución

En este caso el problema consiste en encontrar patrones o características que permitan agrupar los datos (agrupar los perros según su raza, para saber cuántas hay). Por tanto, al no saber a priori cómo se clasifican los datos, el aprendizaje no supervisado sería la mejor opción para descubrir como se agrupan éstos.

En este caso, $\mathcal{X}$ son los datos de los que dispondríamos (las fotos de los perros), $\mathcal{Y}$ sería desconocido ya que no sabemos qué clases hay (no sabemos las razas de perros) y f sería una función de distribución condicional que se quiere aprender con tal de intentar agrupar los datos.

Ejercicio 2

¿Cuáles de los siguientes problemas son más adecuados para una aproximación por aprendizaje y cuáles más adecuados para una aproximación por diseño? Justificar la decisión.

a) Determinar si un vertebrado es mamífero, reptil, ave, anfibio o pez.

Solución

Este problema parece ser más adecuado para el diseño, ya que si se conocen qué características diferencian a los distintos animales, no hace falta aprender nada, solo aplicarlas. Además, por lo general, el problema suele ser bien conocido, con lo cuál la mayoría de características son conocidas, y solo haría falta ajustar unos pocos parámetros para distinguir ciertos casos.

b) Determinar si se debe aplicar una campaña de vacunación contra una enfermedad.

Solución

Este problema parece que puede ser aproximado mejor por diseño que por aprendizaje, ya que es un problema conocido, que es la aplicación de una campaña de vacunas, y solo queremos ajustar algún parámetro, como por ejemplo sería el umbral de personas enfermas desde el que se aplicaría. No haría falta aprender todo el modelo, solo ajustar ese dato. cumplirse una condición que se aplique.

c) Determinar perfiles de consumidor en una cadena de supermercados.

Solución

Para este porblema lo mejor es el aprendizaje, el aprendizaje no supervisado en concreto. No conocemos a priori cuántos perfiles hay y como distinguirlos, pero podemos aplicar alguna técnica de aprendizaje no supervisado con el objetivo de encontrar patrones que permitan distinguir unos perfiles de otros y ver a cuál pertenece un individuo.

d) Determinar el estado anímico de una persona a partir de una foto de su cara.

Solución

La mejor aproximación que se puede seguir en este caso es el aprendizaje, ya que como tal no conocemos exactamente qué detalles de una expresión facial determinan el estado anímico. Si las supiésemos, podríamos simplemente codificar el diseño de éstas, pero como no las sabemos, optaremos por aprender de los datos. Se puede seguir alguna técnica de aprendizaje supervisado o no supervisado para determinar dichos detalles.

e) Determinar el ciclo óptimo para las luces de los semáforos en un cruce con mucho tráfico.

Solución

En este caso, la mejor aproximación que podemos seguir es el aprendizaje, y más concretamente, el aprendizaje por refuerzo. Esto se debe a que se puede construir un simulador donde entrenar un semáforo mediante aprendizaje por refuerzo para que aprenda cuál sería el ciclo óptimo de luces para un determinado cruce con mucho tráfico. Después, se podría trasladar todo lo aprendido al semáforo.

Ejercicio 3

Construir un problema de *aprendizaje desde datos* para un problema de clasificación de fruta en una explotación agraria que produce mangos, papayas y guayabas. Identificar los siguientes elementos formales $\mathcal{X}, \mathcal{Y}, \mathcal{D}, f$ del problema. Dar una descripción de los mismos que pueda ser usada por un computador. ¿Considera que en este problema estamos ante un caso de etiquetas con ruido o sin ruido? Justificar las respuestas.

Solución

Vamos a suponer que nos encontramos ante un problema de clasificación, y por tanto, de aprendizaje supervisado. Para construir nuestro modelo podemos considerar los siguientes elementos:

- $\mathcal{X}$ sería el vector de características de las frutas. Podríamos considerar características tales como el **color**, la **forma**, el **tamaño** y la **textura**.
  - El **color** se podría codificar como una categoría, de tal forma que solo pudiese tomar un valor, como por ejemplo 0 para el verde, 1 para el amarillo y 2 para el verde.
  - La **forma**, al igual que el color, podría tomar un valor categórico, siendo 0 redonda y 1 ovalada.
  - El **tamaño** puede ser también una variable categórica, tomando los valores 0 para pequeño y 1 para grande.
  - La **textura** también puede verse como una variable categórica, pudiendo tomar los valores 0 para lisa y 1 para granulada.
- $\mathcal{Y}$ serían los valores de las etiquetas. Podríamos tener 0 para el **mango**, 1 para la **papaya** y 2 para la **guayaba**.
- $\mathcal{D}$ podría ser en este caso un conjunto de vectores de características con sus correspondientes etiquetas, es decir una muestra, con la cuál pordríamos

entrenar nuestro modelo. Es muy importante que sea una muestra independiente (un elemento no condiciona a los otros) e idénticamente distribuida (cada elemento de la muestra tenga la misma probabilidad).

- f sería nuestra función objetivo, una función desconocida que permitiese asignar las etiquetas a nuestras entradas, es decir, que $f: \mathcal{X} \mapsto \mathcal{Y}$.

En este caso podríamos encontrarnos ante un caso de etiquetas con ruido. Por ejemplo, puede que debido a factores que no hayamos considerado a la hora de establecer las características usadas en $\mathcal{X}$ nos encontremos con que hayan dos frutas con las mismas características, pero que sin embargo luego se hayan clasificado en distintas clases (como puede ser que en algún caso haya habido algúna anomalía durante el crecimiento de una de las frutas y haga que tenga características similares a las de una fruta de la otra clase).

Ejercicio 4

Suponga una matriz cuadrada A que admita la descomposición $\mathrm{A} = \mathrm{X}^{\mathrm{T}}\mathrm{X}$ para alguna matriz X de números reales. Establezca una relación entre los valores singulares de las matriz A y los valores singulares de X.

Solución

Vamos a partir de que la matriz X puede descomponerse en valores singulares de la forma:

$$X = UDV^T \tag{1}$$

donde encontramos que:

- U es una matriz ortogonal, y por tanto, $U^{-1} = U^T$.
- D es una matriz diagonal que contiene los valores singulares de X en su diagonal principal ordenados de mayor a menor.
- V es una matriz ortogonal, de forma que $V^{-1} = V^T$

Al sustituir los valores de X en la descomposición original por la descomposición mostrada en (1), obtenemos que:

$$A = X^T X = (UDV^T)^T(UDV^T) = VDU^TUDV^T = VDDV^T = VD^2V^T \quad (2)$$

Por tanto, al haber supuesto que la matriz A se podía descomponer en $X^T X$, podemos suponer que el resultado obtenido en (2) se corresponde con la descomposición en valores singulares de A. Como hemos partido sustituyendo X por su descomposición en valores singulares, y sabiendo que los valores propios de la matriz X están contenidos en D, entonces podemos decir que los valores propios de A son los de X al cuadrado.

Ejercicio 5

Sean $\mathbf{x}$ e $\mathbf{y}$ dos vectores de características de dimensión $M \times 1$. La expresión

$$\mathrm{cov}(\mathbf{x}, \mathbf{y}) = \frac{1}{M} \sum_{i=1}^{M} (x_i - \overline{x})(y_i - \overline{y})$$

define la covarianza entre dichos vectores, donde $\overline{z}$ representa el valor medio de los elementos de $\mathbf{z}$. Considere ahora una matriz X cuyas columnas representan vectores de características. La matriz de covarianzas asociada a la matriz $\mathrm{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$ es el conjunto de covarianzas definidas por cada dos de sus vectores columnas. Es decir,

$$\mathrm{cov}(\mathbf{X}) = \begin{pmatrix} \mathrm{cov}(\mathbf{x}_1, \mathbf{x}_1) & \mathrm{cov}(\mathbf{x}_1, \mathbf{x}_2) & \cdots & \mathrm{cov}(\mathbf{x}_1, \mathbf{x}_N) \\ \mathrm{cov}(\mathbf{x}_2, \mathbf{x}_1) & \mathrm{cov}(\mathbf{x}_2, \mathbf{x}_2) & \cdots & \mathrm{cov}(\mathbf{x}_2, \mathbf{x}_N) \\ \cdots & \cdots & \cdots & \cdots \\ \mathrm{cov}(\mathbf{x}_N, \mathbf{x}_1) & \mathrm{cov}(\mathbf{x}_N, \mathbf{x}_2) & \cdots & \mathrm{cov}(\mathbf{x}_N, \mathbf{x}_N) \end{pmatrix} \quad (3)$$

Sea $\mathbf{1}^T = (1, 1, \ldots, 1)$ un vector $M \times 1$ de unos. Mostrar que representan las siguientes expresiones:

a) $E1 = \mathbf{1}\mathbf{1}^T\mathrm{X}$

Solución

Sabiendo que $\mathbf{1}$ es un vector $M \times 1$, $\mathbf{1}^T$ es un vector $1 \times M$ y que X es una matriz $M \times N$, podemos aplicar la propiedad asociativa para multiplicar $\mathbf{1}\mathbf{1}^T$, con lo cuál obtendríamos una matriz $M \times M$ de unos. Por tanto, al multiplicar ahora la matriz de unos por X, como éstas tienen dimensiones $M \times M$ y $M \times N$ respectivamente, obtenemos una matriz $M \times N$ en la que todos los elementos de una columna son la suma de los elementos de esa columna. Es decir, para la columna j-ésima de la matriz resultado (de forma que $j \in [1, 2, \ldots, N]$), el elemento i-ésimo de esa columna (de manera que $i \in [1, 2, \ldots, M]$), sería la suma de todos los elementos de la columna j-ésima de la matriz X original.

b) $E2 = (\mathrm{X} - \frac{1}{M}E1)^T(\mathrm{X} - \frac{1}{M}E1)$

Solución

Si comenzamos a operar dentro de los paréntesis, podemos ver que la primera operación que podemos realizar es el producto de $E1$ por un escalar.

Como sabemos de antes, $E1$ es una matriz en la que todos los elementos de una columna son la suma de todos los elementos de la columna correspondiente en X. Al realizar el producto por un escalar, en este caso $\frac{1}{M}$, realmente estamos calculando, para cada elemento de la matriz, la media, ya que como se ha dicho antes, cada elemento de una columna es el sumatorio de los M elementos de la misma columna de X. Con lo cuál, ahora cada elemento de una columna contendrá la media de la suma de los elementos de la misma columna de X. Llamémos a esta matriz $\overline{X}$.

La siguiente operación que podemos realizar, aun dentro de los paréntesis, es $\mathrm{X}-\overline{X}$ (lo que se corresponde con $\mathrm{X}-\frac{1}{M}E1$). Con esto, lo que obtenemos es la diferencia de cada elemento de X con respecto a la media, es decir su desviación con respecto a la media. Esto se puede ver de la siguiente forma:

$$
\begin{aligned}
X - \overline{X} &= \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1N} \\ x_{21} & x_{22} & \cdots & x_{2N} \\ \cdots & \cdots & \cdots & \cdots \\ x_{M1} & x_{M2} & \cdots & x_{MN} \end{pmatrix} - \begin{pmatrix} \overline{x}_1 & \overline{x}_2 & \cdots & \overline{x}_N \\ \overline{x}_1 & \overline{x}_2 & \cdots & \overline{x}_N \\ \cdots & \cdots & \cdots & \cdots \\ \overline{x}_1 & \overline{x}_2 & \cdots & \overline{x}_N \end{pmatrix} \\
&= \begin{pmatrix} x_{11} - \overline{x}_1 & x_{12} - \overline{x}_2 & \cdots & x_{1N} - \overline{x}_N \\ x_{21} - \overline{x}_1 & x_{22} - \overline{x}_2 & \cdots & x_{2N} - \overline{x}_N \\ \cdots & \cdots & \cdots & \cdots \\ x_{M1} - \overline{x}_1 & x_{M2} - \overline{x}_2 & \cdots & x_{MN} - \overline{x}_N \end{pmatrix}
\end{aligned} \tag{4}
$$

Habiendo calculado esto, ahora solo nos queda calcular el producto. La traspuesta de la matriz que se ha obtenido anteriormente es la siguiente (llamemos

X_{dev} a esta matriz, para darle un nombre):

$$X_{dev}^T = \begin{pmatrix} x_{11} - \overline{x}_1 & x_{21} - \overline{x}_1 & \cdots & x_{M1} - \overline{x}_1 \\ x_{12} - \overline{x}_2 & x_{22} - \overline{x}_2 & \cdots & x_{M2} - \overline{x}_2 \\ \cdots & \cdots & \cdots & \cdots \\ x_{1N} - \overline{x}_N & x_{2N} - \overline{x}_N & \cdots & x_{MN} - \overline{x}_N \end{pmatrix} \tag{5}$$

Al realizar la multiplicación ($X_{dev}^T X_{dev}$), lo que obtenemos en realidad no es nada más ni nada menos que una expresión parecida a la covarianza que se puede ver en (3). Es decir, al multiplicar cada fila de X_{dev} por cada columna de X, lo que se obtiene es una sumatoria de $(x_{ij} - x_j)(x_{ji} - x_j)$, para $i \in [1, 2, \ldots, M]$ y $j \in [1, 2, \ldots, N]$ (sabiendo que la matriz X es $M \times N$). Lo único que pasa es que cada elemento está multiplicado por M, ya que no se ha realizado en ningún momento la división. Por tanto, como M multiplica a cada elemento, se puede sacar fuera de esta expresión y dejarlo como un producto de matriz por escalar, pudiendo ahora sí expresar el resultado en función de la covarianza. Por tanto, obtenemos que el valor de $E2$ es el siguiente:

$$E2 = M\mathrm{cov}(\mathrm{X}) \tag{6}$$

Ejercicio 6

Considerar la matriz **hat** definida en regresión, $\hat{\mathrm{H}} = \mathrm{X}(\mathrm{X}^\mathrm{T}\mathrm{X})^{-1}\mathrm{X}^\mathrm{T}$, donde X es la matriz de observaciones de dimensión $N \times (d+1)$, y $\mathrm{X}^\mathrm{T}\mathrm{X}$ es invertible. Justificar las respuestas.

a) ¿Que representa la matriz $\hat{\mathrm{H}}$ en un modelo de regresión?

Solución

b) Identifique la propiedad más relevante de dicha matriz en relación con regresión lineal.

Solución

La propiedad más importante de esta matriz es la idempotencia. Es decir, se da que $\hat{\mathrm{H}}^2 = \hat{\mathrm{H}}$.

Ejercicio 7

La regla de adaptación de los pesos del Perceptron ($\mathbf{w}_{new} = \mathbf{w}_{old} + y\mathbf{x}$) tiene la interesante propiedad de que mueve el vector de pesos en la dirección adecuada para clasificar $\mathbf{x}$ de forma correcta. Suponga el vector de pesos $\mathbf{w}$ de un modelo y un dato $\mathbf{x}(t)$ mal clasificado respecto de dicho modelo. Probar matematicamente que el movimiento de la regla de adaptación de pesos siempre produce un movimiento de $\mathbf{w}$ en la dirección correcta para clasificar bien $\mathbf{x}(t)$.

Ejercicio 8

Ejercicio 9

Ejercicio 10

Referencias

[1] Texto referencia
`https://url.referencia.com`