Introducción y Aprendizaje Paramétrico

Eduardo Morales, Hugo Jair Escalante

INAOE

Introducción

- El área de aprendizaje en general trata de construir programas que mejoren su desempeño automáticamente con la experiencia
- El aprendizaje es posiblemente la característica más distintiva de la inteligencia humana
- Desde el comienzo de las computadoras se cuestionó si serían capaces de aprender
- El entender como aprenden las máquinas nos puede ayudar a entender el aprendizaje humano

Introducción

- Para resolver problemas creamos programas/modelos
- Para algunos problemas es difícil formalizarlos, contar con expertos, es demasiada información, ... $\Rightarrow$ ML
- ML genera automáticamente programas/modelos a partir de datos
- Esto abre una gran cantidad de posibles aplicaciones

Clasificaciones

Varios autores clasifican a los sistemas de aprendizaje de diferentes formas:

- Por el esquema matemático subyacente
- Por la naturaleza de los datos
- Por las tareas que resuelven
- Por las suposiciones sobre los modelos

Esquema Matemático Subyacente

1. Modelos geométricos: Los ejemplos definen un espacio de instancias sobre el cual se pueden construir modelos geométricos, e.g., calculando distancias, buscando hiperplanos, encontrando prototípos, etc.

- Normalmente los atributos son numéricos, por lo que se pueden utilizar conceptos geométricos como líneas, planos y distancias, y se pueden hacer transformaciones lineales y usar diferentes medidas de distancia
- Algunos ejemplos con: Clasificadores lineales, vecinos más cercanos, k-means y clusterin en general, SVMs y clasificaciópn basada en kernesl, clasificación basada en prototipos, etc.

Esquema Matemático Subyacente

2. Modelos probabilistas: En aprendizaje queremos saber cuál es la mejor hipótesis (más probable) dados los datos

- Si $P(D)$ = probabilidad *a priori* de los datos (i.e., cuáles datos son más probables que otros) y $P(D \mid h)$ = probabilidad de los datos dada una hipótesis, lo que queremos estimar es: $P(h \mid D)$, la probabilidad posterior de h dados los datos.
- Esto lo podemos estimar con Bayes.

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

Esquema Matemático Subyacente

- Para estimar la hipótesis más probable o MAP (*maximum a posteriori hypothesis*):

$$\begin{aligned} h_{MAP} &= argmax_{h \in H} \left(P(h \mid D) \right) \\ &= argmax_{h \in H} \left(\frac{P(D|h)P(h)}{P(D)} \right) \\ &\approx argmax_{h \in H} \left(P(D \mid h)P(h) \right) \end{aligned}$$

  Ya que $P(D)$ es una constante independiente de h.

- Si suponemos que las hipótesis son igualmente probables, nos queda la hipótesis de máxima verosimilitud o ML (*maximum likelihood*):

$$h_{ML} = argmax_{h \in H} \left(P(D \mid h) \right)$$

- Aquí se manejan conceptos de probabilidad a priori, máxima verosimilitud, teorema de Bayes, etc.

Esquema Matemático Subyacente

3. Modelos lógicos: Modelos que pueden expresarse desde un punto de vista lógico, incluyendo conjunciones, disjunciones, negación, etc.

- También se les conoce como declarativos, se prestan para dar explicaciones y manejan conceptos como completo y consistente
- Por ejemplo, reglas de clasificación, árboles de decisión, ILP, patrones frecuentes, subgroup discovery, etc.

Naturaleza de los datos

Esta es una de las clasificaciones más usadas:

1. Supervisado: Se tienen datos (X) con una etiqueta (clase) asociada (Y) y se busca encontrar un modelo que dada una instancia de X prediga la etiqueta. Son tareas de clasificación y regresión y se usan conceptos como *overfitting*
2. No supervisado: En este caso no se tienen etiquetas asociadas y se busca encontrar una estructura inherente en los datos, organizándolos generalmente por similaridad o relaciones entre variables
3. Aprendizaje por Refuerzo: Se aprende cómo mapear situaciones a acciones para resolver un problema de decisión secuencial, mediante un proceso iterativo de exploración en el ambiente.

Tareas de Aprendizaje

- Descripción: Se obtienen descripciones de los datos, se realizan resúmenes, se encuentran ejemplos prototípicos, etc.
- Predicción: Se realizan tareas de clasificación (etiquetas discretas) y estimación o regresión (etiquetas contínuas)
- Segmentación: Se separan los datos en subgrupos o clases
- Análisis de dependencias: Se encuentran dependencias entre valores de atributos
- Detección de desviaciones, casos extremos o anomalías
- Control: Aprender qué acción tomar en cada estado
- Optimización y búsqueda

Suposiciones sobre los modelos

- Paramétricos: El modelo resume los datos con un conjunto de parámetros finitos
- No paramétricos: No hacen en general suposiciones fuertes sobre la función o modelo que se quiere encontrar

Modelos Paramétricos

Estos algoritmos involucran dos pasos:

1. Seleccionar la forma de la función
2. Aprender los valores de los coeficientes de la función a partir de los datos

Modelos Paramétricos

Las funciones pueden ser muy variadas, por ejemplo:

- Funciones lineales
- Regresiones logísticas
- Perceptrones
- Naïve Bayes
- Redes neuronales sencillas
- ...

Modelos Paramétricos

Ventajas:

- Simples: Son más fáciles de entender y de interpretar sus resultados
- Velocidad: Se aprenden rápidamente
- Datos: Requieren en general menos datos

Desventajas:

- Restrictivos: Al seleccionar un tipo de función particular se restringe lo que se puede aprender
- Complejidad limitada: En general son adecuados para problemas más sencillos
- Ajuste: Es probable que el modelo seleccionado no ajuste adecuadamente la función subyacente

Modelos No Paramétricos

- No hacen suposiciones acerca de la forma de la función, sino que se determina con los datos
- Algunos ejemplos son:
  - k-vecinos más cercanos
  - Árboles de decisión
  - SVM
  - Aprendizaje Bayesiano
  - ...

Modelos No Paramétricos

Ventajas:

- Flexibilidad: Capaces de ajustar una gran cantidad de funciones
- Poder: No hacen grandes suposiciones sobre los modelos
- Desempeño: Pueden obtener mejores desempeños

Desventajas:

- Datos: Requieren una gran cantidad de datos
- Velocidad: En general se tardan en aprender
- Ajuste: Son propensos a realizar sobre-ajustes

Modelos Paramétricos

- Un estadístico es cualquier valor calculado de alguna muestra
- Empezaremos suponiendo que conocemos la distribución (e.g., gaussiana)
- La ventaja de los modelos paramétricos es que en general dependen de pocos parámetros (e.g., media y varianza)
- La forma que se usa para estimar los parámetros es la de máxima verosimilitud (*maximum likelihood*)

Estimación de Máxima Verosimilitud

- Suponemos una muestra i.i.d. (independiente e identicamente distribuida) $X = \{x_t\}_{t=1}^{N}$
- Suponemos que x_t es una instancia tomada de una familia de distribución conocida, $p(x|\Theta)$, definida por parámetros Θ
- Queremos estimar Θ tal que el muestreo x de $p(x|\Theta)$ sea lo más probable posible

Estimación de Máxima Verosimilitud

- Como x_t es independiente, la verosimilitud de los parámetros dada la muestra X, se calcula como el producto de las verosimilitudes individuales

$$l(\Theta|X) \equiv p(X|\Theta) = \prod_{t=1}^{N} p(x_t|\Theta)$$

- Para obtener la estimación máxima, podemos sacar el logaritmo y convertirlo en una sumatoria

$$\mathcal{L}(\Theta|X) \equiv \log l(X|\Theta) = \sum_{t=1}^{N} \log p(x_t|\Theta)$$

- Para 2 clases, podemos usar una distribución Bernoulli, para N, una gaussiana multinomial

Distribución Bernoulli

- La probabilidad de que ocurra el evento $(X = 1)$ es: p y que no ocurra $(X = 0)$ es: $1 - p$

$$P(x) = p^x(1-p)^{1-x}, x \in \{0, 1\}$$

- El valor esperado y varianza de X son:

$$E[X] = \sum_x xp(x) = 1 \cdot p + 0 \cdot (1-p) = p$$

$$Var(X) = \sum_x (x - E[X])^2 p(x) = p(1-p)$$

Distribución Bernoulli

- El único parámetro que se tiene es p, y queremos calcular un estimado $\hat{p}$
- El *log likelihood* es:

$$\mathcal{L}(p|X) = \log \prod_{t=1}^{N} p^{x_t}(1-p)^{1-x_t}$$

$$= \sum_t x_t \log p + (N - \sum_t x_t) \log(1-p)$$

- Para maximizar, derivamos con respecto a p, $d\mathcal{L}/dp = 0$

$$\hat{p} = \frac{\sum_t x_t}{N}$$

- Que es lo esperado

Distribución Normal

- $Var(x) \equiv \sigma^2, E[X] = \mu$

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

- El *log likelihood* es:

$$\mathcal{L}(\mu, \sigma|X) = -\frac{N}{2}\log(2\pi) - N\log\sigma - \frac{\sum_t (x_t - \mu)^2}{2\sigma^2}$$

- Tomando las derivadas parciales de cada argumento e igualando a cero, tenemos:

$$m = \frac{\sum_t x_t}{N}$$

$$s^2 = \frac{\sum_t (x_t - m)^2}{N}$$

Clasificación Paramétrica

- Para hacer una clasificación Bayesiana:

$$p(C_i|x) = \frac{p(x|C_i)p(C_i)}{p(x)}$$

- Dado que el denominador es igual, la función de discriminación es:

$$g_i(x) = p(x|C_i)p(C_i)$$

- O de forma equivalente:

$$g_i(x) = \log p(x|C_i) + \log p(C_i)$$

Clasificación Paramétrica

- Suponiendo que $p(x|C_i)$ es gaussiana:

$$p(x|C_i) = \frac{1}{\sqrt{2\pi}\sigma_i} exp\left[-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right]$$

- La función de discriminación es entonces:

$$g_i(x) = -\frac{1}{2} log(2\pi) - log\sigma_i - \frac{(x-\mu_i)^2}{2\sigma_i^2} + logp(C_i)$$

- Podemos estimar la media, desviación estandar y $P(C_i)$ de los datos, con los que nos queda:

$$g_i(x) = -\frac{1}{2} log(2\pi) - logs_i - \frac{(x-m_i)^2}{2s_i^2} + logp(C_i)$$

Clasificación Paramétrica

- El primer término es constante y si suponemos que las probabilidades de la clase son iguales y que las varianzas también, nos queda:

$$g_i(x) = -(x - m_i)^2$$

- Por lo que asignamos la clase al elemento que esté más cercano a la media:
- C_i si $|x - m_i| = min_k|x - m_k|$
- Con dos clases, el punto medio es el umbral de decisión: $g_1(x) = g_2(x)$

$$(x - m_i)^2 = (x - m_2)^2$$

$$x = \frac{m_1 + m_2}{2}$$

Regresión

- En regresión la salida o variable dependiente es función de las entradas o variables independientes $r = f(x) + \epsilon$
- $f(x)$ es una función desconocida que queremos estimar con $g(x|\Theta)$, definida por un conjunto de parámetros Θ
- Si suponemos que ϵ es ruido con una distribución gaussiana con media cero y varianza constante ($\epsilon \sim \mathcal{N}(0, \sigma^2)$), poniendo nuestro estimador $g(\cdot)$ en lugar de $f(\cdot)$:

$$p(r|x) \sim \mathcal{N}(g(x|\Theta), \sigma^2)$$

Regresión

- De nuevo queremos aprender los parámetros Θ usando máxima verosimilitud
- Lo que tenemos es un conjunto de datos (x, r) dados de una cierta densidad de distribución $p(x, r)$, que podemos escribir como:

$$p(x, r) = p(r|x)p(x)$$

- Y el logaritmo de su verosimilitud es:

$$\mathcal{L}(\Theta|X) = \log \prod_{t=1}^{N} p(x_t, r_t)$$

$$= \log \prod_{t=1}^{N} p(r_t|x_t) + \log \prod_{t=1}^{N} p(x_t)$$

Regresión

- Ignorando el segundo término que no depende de nuestro estimador:

$$\mathcal{L}(\Theta|X) = \log \prod_{t=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} exp\left[-\frac{(r_t - g(x_t|\Theta))^2}{2\sigma^2}\right]$$

$$= \log\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N exp\left[-\frac{1}{2\sigma^2}\sum_{t=1}^{N}(r_t - g(x_t|\Theta))^2\right]$$

$$= -N\log(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2}\sum_{t=1}^{N}(r_t - g(x_t|\Theta))^2$$

- El primer elemento es independiente de los parámetros al igual que $1/\sigma^2$, lo que nos da la función de error más frecuentemente usada:

$$= -\frac{1}{2}\sum_{t=1}^{N}(r_t - g(x_t|\Theta))^2$$

Regresión

- En regresión lineal, lo que tenemos es: $g(x_t|w_1, w_0) = w_1 x_t + w_0$
- Si sacamos la derivada de la función de error con respecto a w_1 y w_0 nos queda

$$\sum_t r_t = N w_0 + w_1 \sum_t x_t$$

$$\sum_t r_t x_t = w_0 \sum_t x_t + w_1 \sum_t (x_t)^2$$

- Lo que se puede re-escribir en forma matricial como: $Aw = y$ donde:

$$A = \begin{bmatrix} N & \sum_t x_t \\ \sum_t x_t & \sum_t (x_t)^2 \end{bmatrix}$$

$$w = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}; y = \begin{bmatrix} \sum_t r_t \\ \sum_t r_t x_t \end{bmatrix}$$

Regresión

- Lo que se resuelve como $w = A^{-1}y$
- Esto mismo se puede extender a regresiones polinomiales, transformandolo a una forma $Ax = y$
- Se hace más o menos lo mismo para problemas multivariables
- También es importante determinar el orden de los polinomios. Entre más alto, tienen más varianza con pequeños cambios en los datos pero puede ajustar mejor
- Se tiene un balance entre el sesgo y la varianza (*bias/variance*)

Entropía Cruzada

- Otra aplicación común es querer aprender una función de probabilidad sobre un función que tiene dos posibles resultados (e.g., 0 o 1)
- Si los datos (D) son: $D = \{(x_1, d_1), \ldots, (x_m, d_m)\}$, donde d_i es el valor observado (0 o 1) de $f(x_i)$, y suponiendo que los datos son independientes:

$$P(D \mid h) = \prod_{i=1}^{m} P(x_i, d_i \mid h)$$

- Si x_i es independiente de h

$$P(D \mid h) = \prod_{i=1}^{m} P(d_i \mid h, x_i) P(x_i)$$

Entropía Cruzada

- Como h es la probabilidad de la función meta $P(d_i = 1 \mid h, x_i) = h(x_i)$, y en general:

$$P(d_i \mid h, x_i) = \begin{cases} h(x_i) & \text{si } d_i = 1 \\ 1 - h(x_i) & \text{si } d_i = 0 \end{cases}$$

- Esto mismo lo podemos expresar como:

$$P(d_i \mid h, x_i) = h(x_i)^{d_i}(1 - h(x_i))^{1-d_i}$$

- Por lo que:

$$P(D \mid h) = \prod_{i=1}^{m} h(x_i)^{d_i}(1 - h(x_i))^{1-d_i} P(x_i)$$

Entropía Cruzada

- La máxima verosimilitud es entonces:

$$h_{ML} = argmax_{h \in H} \left(\prod_{i=1}^{m} h(x_i)^{d_i} (1 - h(x_i))^{1-d_i} P(x_i) \right)$$

- Al eliminar el último término (que no depende de h), tenemos:

$$h_{ML} = argmax_{h \in H} \left(\prod_{i=1}^{m} h(x_i)^{d_i} (1 - h(x_i))^{1-d_i} \right)$$

- Lo cual es una generalización de la distribución Binomial (e.g., describe la probabilidad de que al tirar m monedas se tenga $(d_1, \ldots, d_m)$ resultados suponiendo que cada moneda tiene probabilidad $h(x_i)$ de salir sol)

Entropía Cruzada

- En la descripción de la distribución binomial se supone que todas las monedas tienen la misma probabilidad de que salga sol.
- Trabajando (de nuevo) con el logaritmo:

$$h_{ML} = argmax_{h \in H} \left(\sum_{i=1}^{m} d_i ln(h(x_i)) + (1 - d_i) ln(1 - h(x_i)) \right)$$

- Lo cual, por su parecido con la medida de entropía, también se llama a su negativo, entropía cruzada (*cross entropy*).

Funciones de Pérdida

- Las funciones de pérdida se utilizan para optimizar un modelo (minimizar pérdida)
- Dos de las funciones de pérdida más usadas en ML son:
  - Error cuadrático medio (MSE o L2): Regresión

$$MSE = -\frac{1}{2}\sum_{t=1}^{N}(y_i - \hat{y}_i)^2$$

  - Entropía cruzada: Clasificación Para 2 clases:

$$CE(y, p) = -y\log(p) - (1-y)\log(1-p)$$

    Para m clases:

$$CE(y, p) = -\sum_{c=1}^{m} y_{o,c} log(p_{o,c})$$

    $y =$ indicador binario (0 o 1) y $p =$ probabilidad de la clase dada una observación o.
- Otras: Hinge, Huber, Kullback-Leibler, RMSE, MAE (L1)