Aprendizaje generativo

Reconocimiento de patrones

Gamaliel Moreno

gamalielmch@uaz.edu.mx
http://pds.uaz.edu.mx/

Enero-Julio 2021

Contenido

Aprendizaje discriminador

- Hasta ahora, aprendizaje basado en $p(y|\boldsymbol{x};\boldsymbol{\theta})$
  - Regresión logística: $p(y|\boldsymbol{x};\boldsymbol{\theta}) = h_{\boldsymbol{\theta}}(\boldsymbol{x}) = g(\boldsymbol{\theta}^T\boldsymbol{x})$ con $g(\cdot)$ sigmoidal.
- Concepto actual ha particionado el espacio de características con un borde de decisión
- Clasificación se reduce a evaluar en qué lado del borde de decisión está la entrada
- Algortimos que aprenden $p(y|x)$ directamente se llaman algoritmos discriminadores
- Pueden aprender $h_{\boldsymbol{\theta}}(\boldsymbol{x}) \in \{0, 1\}$

Aprendizaje generativo

- Otra idea: aprender $p(\boldsymbol{x}|y)$ y $p(y)$
- Ejemplo: aprendamos con características de forma/textura modelos para
  - Cáncer benigno
  - Cáncer maligno
- Para cada clase aprendemos un modelo por separado
- Para predicción, deben probarse todos los modelos y se selecciona el más probable
- Este enfoque se denomina aprendizaje generativo.

Análisis Gaussiano discriminador

- Supongamos que $\boldsymbol{x} \in \mathbb{R}^n$ son continuos
- Además supongamos que $p(\boldsymbol{x}|y)$ es guassiano

$$p(\boldsymbol{x}|y) = \mathbf{N}(\mu, \Sigma)$$

donde μ es la media y Σ es la matriz de covarianza

Análisis Gaussiano discriminador

- Supongamos que

$$p(y) = \phi^y(1-\phi)^{1-y}$$

$$p(\mathbf{x}|y=0) = \frac{1}{\sqrt{(2\pi)^n|\mathbf{\Sigma}|}} exp\left(-\frac{1}{2}(\mathbf{x}-\mu_0)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_0)\right)$$

$$p(\mathbf{x}|y=1) = \frac{1}{\sqrt{(2\pi)^n|\mathbf{\Sigma}|}} exp\left(-\frac{1}{2}(\mathbf{x}-\mu_1)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_1)\right)$$

- Buscamos entonces maximizar la verosimilitud

$$\ell(\phi,\mu_0,\mu_1,\mathbf{\Sigma}) = ln \underbrace{\prod_i^m p(\mathbf{x}^{(i)},y^{(i)})}_{\text{verosimilitud conjunta}} = \prod_i^m p(\mathbf{x}^{(i)}|y^{(i)})p(y^{(i)})$$

- Esta verosimilitud conjunta

$$\ell(\phi, \mu_0, \mu_1, \boldsymbol{\Sigma}) = \prod_{i}^{m} p(\boldsymbol{x}^{(i)}|y^{(i)})p(y^{(i)})$$

contrasta con la verosimilitud condicional utilizada en regresión logística

$$\ell(\boldsymbol{\theta}) = ln \prod_{i}^{m} p(y^{(i)}|\boldsymbol{x}^{(i)}; \boldsymbol{\theta})$$

Análisis Gaussiano discriminador

- Maximizando la verosimilitud anterior se obtiene

$$\phi = \frac{1}{m} \sum_{i=1}^{m} 1\{y^{(i)} = 1\}$$

$$\boldsymbol{\mu}_0 = \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 0\} \boldsymbol{x}^{(i)}}{\sum_{i=1}^{m} 1\{y^{(i)} = 0\}}$$

$$\boldsymbol{\mu}_1 = \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 1\} \boldsymbol{x}^{(i)}}{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}}$$

$$\boldsymbol{\Sigma} = \frac{1}{m} \sum_{i=1}^{m} (\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_{y^{(i)}})(\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_{y^{(i)}})^T$$

Predicción

- Observemos que con la regla de Bayes, puede recalcularse:

$$p(y = 1|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|y = 1)p(y)}{p(\boldsymbol{x})}$$

$$p(\boldsymbol{x}) = p(\boldsymbol{x}|y = 0)p(y = 0) + p(\boldsymbol{x}|y = 1)p(y = 1)$$

- sin embargo, $p(\boldsymbol{x})$ usualmente es innecesario pues para la predicción basta con:

$$\arg\max_y p(y|\boldsymbol{x}) = \arg\max_y \frac{p(\boldsymbol{x}|y)p(y)}{p(\boldsymbol{x})} = \arg\max_y p(\boldsymbol{x}|y)p(y)$$

- Si $p(y)$ es uniforme $(p(y = 0) = p(y = 1))$ entonces: $\arg\max_y p(\boldsymbol{x}|y)$

GDA

- Dado el conjunto de entrenamiento $(\boldsymbol{x}^{(i)}, y^{(i)})$
- Calcular con el conjunto los parámetros $\boldsymbol{\mu}_i, \Sigma$ y $p(y)$
- Para predecir probabilidad de y dado un valor de $\boldsymbol{x}$:
  - Calculamos con parámetros $p(\boldsymbol{x}|y=0) = \mathbf{N}(\boldsymbol{\mu}_0, \sigma_0^2)$ y $p(\boldsymbol{x}|y=1) = \mathbf{N}(\boldsymbol{\mu}_1, \sigma_1^2)$
  - con eso calculamos

$$p(y=1|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|y=1)p(y)}{p(\boldsymbol{x})}$$

donde $p(\boldsymbol{x})$ se calcula con

$$p(\boldsymbol{x}) = p(\boldsymbol{x}|y=0)p(y=0) + p(\boldsymbol{x}|y=1)p(y=1)$$

Ventajas y desventajes de algoritmos generativos

- En GDA supusimos $\boldsymbol{x}|y \backsim$ gaussiano
- Eso implica que la distribución a-posteriori $p(y=1|\boldsymbol{x})$ es logística
- Lo contario no es cierto: logístico $\nrightarrow \boldsymbol{x}|y \backsim$ gaussiano
- (por ejemplo, si $\boldsymbol{x}|y \backsim$ Poisson también la probabilidad a-posteriori es logística)
- Eso implica que suposición del GDA es más fuerte
- Si la suposición es cierta, entonces GDA es mejor que la regresión logística
- Si no se sabe qué distribución tiene los datos, entonces la regresión logística es una mejor elección
- GDA funciona a veces mejor con pocos datos
- Regresión logística requiere por lo general más datos

Clasificador bayesiano ingenuo
(Segundo método generativo)

Características

- Representación usa un vector de dimensión igual al número de palabras en el diccionario
- Si el correo-e contiene la iésima palabra del diccionario usamos $x_i = 1$, y caso contrario $x_i = 0$
- Por ejemplo

$$x = \begin{array}{cc} \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} & \begin{array}{c} a \\ ababa \\ ababillarse \\ \vdots \\ compra \\ \vdots \\ zwingliano \end{array} \end{array}$$

Vocabulario

- Conjunto de palabra codificada en el vector de características se llama vocabulario
- Tamaño del vocabulario igual a dimensión de $\boldsymbol{x}$
- Si tenemos un vocabulario de 50,000 palabras, entonces $\boldsymbol{x} \in \{0; 1\}$
- Queremos armar un modelo generativo, así que necesitamos un modelo para $p(\boldsymbol{x}|y)$
- Obviamente no es posible modelar cada $\boldsymbol{x}$ explícitamente con un modelo multinomial, pues tendríamos 2^{50000} posibles configuraciones, lo que implica un vector de configuración de $(2^{50000} - 1)$ dimensiones

Probabilidad conjunta condicional

$$
\begin{aligned}
p(\boldsymbol{x}|y) &= p(x_1, \ldots, x_{50,000}|) \\
&= p(x_1|y)p(x_2|y, x_1) \cdot p(x_{50,000}|y, x_1, \ldots x_{49,999}) \\
&= p(x_1|y)p(x_2|y) \cdots p(x_{50,000}|y) \\
&= \prod_{i=1}^{n} p(x_i|y)
\end{aligned}
$$

- A pesar de que esta suposición es muy fuerte, el método funciona.
- El modelo se parametriza con $\phi_{i|y=1} = p(x_i = 1|y = 1)$, $\phi_{i|y=0} = p(x_i = 1|y = 0)$ y $\phi_y = p(y = 1)$

Máxima verosimilitud

- Si se maximiza $L(\phi_y, \phi_{j|y=0}, \phi_{j|y=1})$ con respecto a los parámetros, se obtiene el estimado de máxima verosimilitud:

$$\phi_{j|y=1} = p(x_j = 1|y = 1) = \frac{\sum_i^m 1\{x_j^{(i)} \wedge y^{(i)} = 1\}}{\sum_i^m \{y^{(i)} = 1\}}$$

$$\phi_{j|y=0} = p(x_j = 1|y = 0) = \frac{\sum_i^m 1\{x_j^{(i)} \wedge y^{(i)} = 0\}}{\sum_i^m \{y^{(i)} = 0\}}$$

$$\phi_y = p(y = 1) = \frac{\sum_i^m 1\{y^{(i)} = 1\}}{m}$$

- Interpretaciones fáciles

Máxima verosimilitud

- Con estos parámetros, para hacer la predicción en un nuevo correo $\boldsymbol{x}$ solo calculamos:

$$p(y=1|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|y=1)p(y=1)}{p(c)}$$

$$= \frac{\left(\prod_{i=1}^{n} p(x_i|y=1)\right) p(y=1)}{\left(\prod_{i=1}^{n} p(x_i|y=1)\right) p(y=1) + \left(\prod_{i=1}^{n} p(x_i|y=0)\right) p(y=0)}$$

- Elegimos la clase que tenga la probabilidad a-posteriori mayor

Caso multinomial

- Desarrollamos el algoritmo de Bayes ingenuo para características de entrada x_i binarias
- Nada impide usar características $x_i \in \{1, 2, \ldots, k_i\}$
- En ese caso modelamos $p(x_i|y)$ con una distribución multinomial en vez de Bernoulli
- En la práctica, en problemas con entras continuas, se obtienen buenos resultados se se discretiza la entrada y se usa el algoritmo de Bayes ingenuo (por ejemplo, si datos no siguen una dsitribución normal multivariada)

Suavizamiento con Laplace

Suavizamiento con laplace

- El algoritmo ingenuo de Bayes funciona en bastantes problemas
- Un cambio simple lo mejora, especialmente para clasificación textual
- Una nueva palabra k que no estuvo en el conjunto de entrenamiento tendrá:

$$\phi_{k|y=1} = \frac{\sum_i^m 1\{x_j^{(i)} \wedge y^{(i)} = 1\}}{\sum_i^m \{y^{(i)} = 1\}} = 0$$

$$\phi_{k|y=0} = \frac{\sum_i^m 1\{x_j^{(i)} \wedge y^{(i)} = 0\}}{\sum_i^m \{y^{(i)} = 0\}} = 0$$

Problema con suposición ingenua

- Como la plabra no es spam no no-spam, la probabilidad de que cualquiera ocurra es cero
- Si queremos decidir qué tipo de correo es uno que contenga la k-ésima palabra se obtiene:

$$\frac{\left(\prod_{i=1}^{n} p(x_i|y=1)\right) p(y=1)}{\left(\prod_{i=1}^{n} p(x_i|y=1)\right) p(y=1) + \left(\prod_{i=1}^{n} p(x_i|y=0)\right) p(y=0)}$$
$$= \frac{0}{0}$$

Suavizamiento de Laplace

- Estadísticamente es mala idea suponer que la probabilidad de un evento es cero solo porque no se ha visto en el conjunto de entrenamiento
- Para m observaciones, estimación de máxima verosimilitud es:

$$\phi_j = \frac{\sum_{i=1}^{m} 1\{x^i = j\}}{m}$$

- Con esta estimación algunos phi_j pueden llegar a ser cero, lo que se evita con el suavizamiento de Laplace, que lo reemplaza con

$$\phi_j = \frac{\sum_{i=1}^{m} 1\{x^i = j\} + 1}{m + k}$$

- k es el número de posibles valores que puede tomar $x^{(i)}$ (en el caso binario k=2)
- Note que aún se cumple $\sum_{j=1}^{k} \phi_j = 1$ y $\phi_j \neq 0$