Regularización y aprendizaje en línea

Lección 16

Dr. Pablo Alvarado Moya

CE5506 Introducción al reconocimiento de patrones
Área de Ingeniería en Computadores
Tecnológico de Costa Rica

II Semestre, 2019

Contenido

Enfoque frecuentista

- Retomemos como ejemplo regresión lineal/logística
- Elegimos parámetros que maximizan la verosimilitud

$$\max_{\underline{\theta}} \prod_i p(y^{(i)}|\underline{\mathbf{x}}^{(i)};\underline{\theta})$$

- Aquí, $\underline{\theta}$ **no** es una variable aleatoria, sino una constante desconocida
- Nuestra tarea es encontrar métodos para estimar ese vector constante desconocido $\underline{\theta}$
- Este es el enfoque usual en estadística **frecuentista**.

Enfoque Bayesiano (1)

- En el enfoque estadístico **Bayesiano** $\underline{\theta}$ también es una variable aleatoria de valor desconocido
- Se especificaría una distribución **a priori** $p(\underline{\theta})$ que expresa las creencias previas sobre los parámetros
  (antes de ver cualquier dato)
- Por ejemplo: $p(\underline{\theta}) \sim \mathcal{N}(\underline{\mu}, \mathbf{\Sigma})$

Enfoque Bayesiano (2)

- Dado el conjunto de entrenamiento $\mathcal{S} = \{(\underline{\mathbf{x}}^{(i)}, y^{(i)}) \mid i = 1 \ldots m\}$ se puede calcular la distribución *a posteriori* de los parámetros:

$$p(\underline{\boldsymbol{\theta}}|\mathcal{S}) = \frac{p(\mathcal{S}|\underline{\boldsymbol{\theta}})p(\underline{\boldsymbol{\theta}})}{p(\mathcal{S})}$$
$$= \frac{\left(\prod_{i=1}^{m} p(y^{(i)}|\underline{\mathbf{x}}^{(i)}, \underline{\boldsymbol{\theta}})\right) p(\underline{\boldsymbol{\theta}})}{\int_{\underline{\boldsymbol{\theta}}} \left(\prod_{i=1}^{m} p(y^{(i)}|\underline{\mathbf{x}}^{(i)}, \underline{\boldsymbol{\theta}}) p(\underline{\boldsymbol{\theta}})\right) \, d\underline{\boldsymbol{\theta}}}$$

(después de ver los datos de entrenamiento)

Eligiendo la verosimilitud

- A $p(y^{(i)}|\underline{\mathbf{x}}^{(i)},\underline{\boldsymbol{\theta}})$ lo determina el modelo que se usa para el aprendizaje
- Con regresión logística Bayesiana, puede elegirse:

$$p\left(y^{(i)}\middle|\underline{\mathbf{x}}^{(i)},\underline{\boldsymbol{\theta}}\right) = h_{\underline{\boldsymbol{\theta}}}\left(\underline{\mathbf{x}}^{(i)}\right)^{y^{(i)}}\left(1-h_{\underline{\boldsymbol{\theta}}}\left(\underline{\mathbf{x}}^{(i)}\right)\right)^{1-y^{(i)}}$$

con

$$h_{\underline{\boldsymbol{\theta}}}(\underline{\mathbf{x}}^{(i)}) = \frac{1}{1+\exp(-\underline{\boldsymbol{\theta}}^{T}\underline{\mathbf{x}}^{(i)})}$$

Predicción

- Cuando nos piden predecir y para un nuevo $\underline{\mathbf{x}}$ calculamos la distribución *a posteriori*

$$p(y|\underline{\mathbf{x}},\mathcal{S}) = \int_{\underline{\boldsymbol{\theta}}} p(y|\underline{\mathbf{x}},\underline{\boldsymbol{\theta}})p(\underline{\boldsymbol{\theta}}|\mathcal{S})\, d\underline{\boldsymbol{\theta}}$$

- Si queremos predecir el valor esperado de y dado $\underline{\mathbf{x}}$, entonces:

$$\mathrm{E}[y|\underline{\mathbf{x}},\mathcal{S}] = \int_{y} yp(y|\underline{\mathbf{x}},\mathcal{S})\, dy$$

- Estos pasos son una predicción completamente Bayesiana
- Calcular la distribución *a posteriori* $p(\underline{\boldsymbol{\theta}}|\mathcal{S})$ es en general difícil, por la necesidad de integrar en $\underline{\boldsymbol{\theta}}$ (usualmente de muchas dimensiones), y no se puede hacer de forma cerrada.

Aproximación de distribución *a posteriori* (1)

- Por lo general aproximaremos la distribución *a posteriori* de $\underline{\boldsymbol{\theta}}$
- Una opción usual es reemplazar la distribución anterior con una estimación con un solo punto.
- El estimado MAP (*maximum a posteriori*) de $\underline{\boldsymbol{\theta}}$ es

$$\underline{\hat{\boldsymbol{\theta}}}_{\text{MAP}} = \arg\max_{\underline{\boldsymbol{\theta}}} p(\underline{\boldsymbol{\theta}}|\mathcal{S}) = \arg\max_{\underline{\boldsymbol{\theta}}} \left(\prod_{i=1}^{m} p(y^{(i)}|\underline{\mathbf{x}}^{(i)}, \underline{\boldsymbol{\theta}}) \right) p(\underline{\boldsymbol{\theta}})$$

  y para predecir se usa $h_{\underline{\hat{\boldsymbol{\theta}}}_{\text{MAP}}}(\underline{\mathbf{x}}) = g(\underline{\hat{\boldsymbol{\theta}}}_{\text{MAP}}^{T} \underline{\mathbf{x}})$
- Esto es muy parecido a máxima verosimilitud, excepto por el término *a priori* $p(\underline{\boldsymbol{\theta}})$

Aproximación de distribución *a posteriori* (2)

- En la práctica se usa $\underline{\boldsymbol{\theta}} \sim \mathcal{N}(0, \tau^2 \mathbf{I}))$
  (por ejemplo, clasificación de texto con 50 000 parámetros)
- Con eso los parámetros $\hat{\underline{\boldsymbol{\theta}}}_{\text{MAP}}$ usualmente tienen menor norma que los elegidos con ML, lo que los hace menos susceptibles a sobreajuste.
- Esto se usa con éxito en clasificación de textos

Función de error equivalente

- Surge la pregunta: si utilizamos MAP con la estrategia anterior, ¿cuál es la función de error que estamos minimizando?
- Resulta que la nueva función de error para regresión lineal es:

$$J(\underline{\boldsymbol{\theta}}) = \sum_{i=1}^{m} \left| y^{(i)} - \underline{\boldsymbol{\theta}}^T \underline{\mathbf{x}}^{(i)} \right|^2 + \lambda \|\underline{\boldsymbol{\theta}}\|_2^2$$

- El término de regularización $\lambda \|\underline{\boldsymbol{\theta}}\|_2^2$ fuerza pequeños valores de los parámetros, lo que se puede interpretar como una simplificación del modelo
- Es usual dejar a θ_0 fuera de la regularización
- Las ecuaciones normales son en este caso:

$$\underline{\boldsymbol{\theta}} = (\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I})^{-1}\mathbf{X}^T\underline{\mathbf{y}}$$

Regresión logística con regularización

- En el caso de regresión logística, regularizamos con

$$J(\underline{\boldsymbol{\theta}}) = \ell(\underline{\boldsymbol{\theta}}) - \lambda \|\underline{\boldsymbol{\theta}}\|_2^2$$
$$= \left(\sum_{i=1}^{m} y^{(i)} \ln(h_{\underline{\boldsymbol{\theta}}}(\underline{\mathbf{x}}^{(i)})) + (1 - y^{(i)}) \ln(1 - h_{\underline{\boldsymbol{\theta}}}(\underline{\mathbf{x}}^{(i)})) \right) - \lambda \|\underline{\boldsymbol{\theta}}\|_2^2$$

con gradiente

$$\nabla_{\underline{\boldsymbol{\theta}}} J(\underline{\boldsymbol{\theta}}) = \left[\sum_{i=1}^{\infty} \left(y^{(i)} - h_{\underline{\boldsymbol{\theta}}}(\underline{\mathbf{x}}^{(i)}) \right) \underline{\mathbf{x}}^{(i)} \right] - 2\lambda \underline{\boldsymbol{\theta}}$$

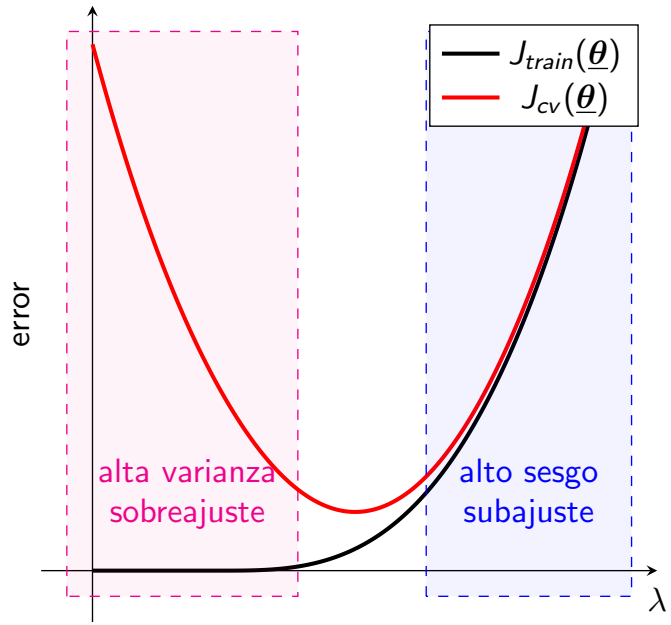
donde $\lambda \|\underline{\boldsymbol{\theta}}\|_2^2$ se resta puesto que aquí **maximizamos** $J(\underline{\boldsymbol{\theta}})$.

Ejemplos

1. Regresión polinomial: `nregress_norm.m`
2. Regresión logística: `rlr.m`

Selección de λ

- En general optimizamos con términos de regularización $\lambda\|\underline{\boldsymbol{\theta}}\|_2^2$ (excluyendo a θ_0).
- La estimación del error para análisis se hace **sin** ese término.



Tipos de aprendizaje

- Hemos trabajado hasta ahora con aprendizaje **fuera de línea** (*off-line*)
- En particular trabajamos con aprendizaje **por lotes** y aprendizaje **secuencial**
- En estos métodos entrenamos primero la hipótesis h con un conjunto de datos y probamos en un conjunto aparte de prueba
- Existe además el aprendizaje **en línea**
- En estos métodos el algoritmo debe predecir contínuamente mientras aprende.

Definición de aprendizaje en línea

- Dado un conjunto de datos

$$(\underline{\mathbf{x}}^{(1)},y^{(1)}),(\underline{\mathbf{x}}^{(2)},y^{(2)}),\ldots,(\underline{\mathbf{x}}^{(m)},y^{(m)})$$

  que se producen en exactamente ese orden.
- El algoritmo primero recibe $\underline{\mathbf{x}}^{(1)}$, y predice $\hat{y}^{(1)}$
- Después de eso se le revela al algoritmo el valor verdadero $y^{(1)}$
- Con $y^{(1)}$ el algoritmo aprende
- Ahora se le presenta al algoritmo $\underline{\mathbf{x}}^{(2)}$, y predice $\hat{y}^{(2)}$.
- Con $y^{(2)}$ algoritmo aprende.
- Proceso se repite hasta m-ésima iteración
- Nos interesa el **error en-línea total**:

$$\varepsilon = \sum_{i=1}^{m} 1\left\{\hat{y}^{(i)} \neq y^{(i)}\right\}$$

Perceptron en-línea

- Se podría utilizar cualquier algoritmo de los anteriories usando todos los datos previos
- Aprendizaje secuencial se adapta casi directamente al caso:
- Por ejemplo, el perceptron

  Inicialice $\underline{\boldsymbol{\theta}} = \underline{\mathbf{0}}$, $i = 1$

  repeat

  $\underline{\boldsymbol{\theta}} \leftarrow \underline{\boldsymbol{\theta}} + \alpha(y^{(i)} - h(\underline{\mathbf{x}}^{(i)}))\underline{\mathbf{x}}^{(i)}$

  $i \leftarrow i + 1$

  until *convergencia*
- Se ha demostrado que esto siempre converge para datos separables con un margen geométrico γ.
- Ver notas `cs229-notes6.pdf` de Andrew Ng

Resumen

Este documento ha sido elaborado con software libre incluyendo LaTeX, Beamer, GNUPlot, GNU/Octave, XFig, Inkscape, GNU-Make y Subversion en GNU/Linux