Universidad Técnica Federico Santa María
Departamento de Informática

Control 3 - Máquinas de Aprendizaje (INF-578)

Semestre II 2017 - Viernes 22.12.17

PAUTA

1. (10 puntos) Asumiendo una función de pérdida cuadrática, escriba la descomposición sesgo-varianza del error de un ensamblado y explique su significado.

Respuesta: Si denotemos como $f_S(\mathbf{x})$ a la hipótesis obtenida a partir de un conjunto de entrenamiento S, como $\mathbb{E}_S$ al valor esperado con respecto a la distribución conjunta de la muestra S, y como $f_0(\mathbf{x})$ a la hipótesis que minimiza el error de predicción en el espacio de hipótesis bajo consideración, tenemos que

$$\begin{aligned}\mathbb{E}_S(f_0(\mathbf{x}) - f_S(\mathbf{x}))^2 &= \mathbb{E}_S(f_0(\mathbf{x}) - \mathbb{E}_S f_S(\mathbf{x}) + \mathbb{E}_S f_S(\mathbf{x}) - f_S(\mathbf{x}))^2 \\ &= \mathbb{E}_S(f_0(\mathbf{x}) - \mathbb{E}_S f_S(\mathbf{x}))^2 + \mathbb{E}_S(\mathbb{E}_S f_S(\mathbf{x}) - f_S(\mathbf{x}))^2 \\ &+ 2\mathbb{E}_S(f_0(\mathbf{x}) - \mathbb{E}_S f_S(\mathbf{x}))(\mathbb{E}_S f_S(\mathbf{x}) - f_S(\mathbf{x})) \\ &= (f_0(\mathbf{x}) - \mathbb{E}_S f_S(\mathbf{x}))^2 + \mathbb{E}_S(\mathbb{E}_S f_S(\mathbf{x}) - f_S(\mathbf{x}))^2 \ , \end{aligned} \tag{1}$$

ya que

$$2\mathbb{E}_S(f_0(\mathbf{x}) - \mathbb{E}_S f_S(\mathbf{x}))(\mathbb{E}_S f_S(\mathbf{x}) - f_S(\mathbf{x})) = (f_0(\mathbf{x}) - \mathbb{E}_S f_S(\mathbf{x}))(\mathbb{E}_S f_S(\mathbf{x}) - \mathbb{E}_S f_S(\mathbf{x})) = 0 \ . \tag{2}$$

La ecuación (1) se suele escribir como

$$\mathbb{E}_S(f_0(\mathbf{x}) - f_S(\mathbf{x}))^2 = \text{Bias}^2(f_S(\mathbf{x})) + \text{Var}(f_S(\mathbf{x})) \ , \tag{3}$$

definiendo, para cualquier hipótesis $f_S(\mathbf{x})$ obtenida a partir de S

$$\text{Bias}(f_S(\mathbf{x})) = (f_0(\mathbf{x}) - \mathbb{E}_S f_S(\mathbf{x})) \tag{4}$$

$$\text{Var}(f_S(\mathbf{x})) = \mathbb{E}_S(f_S(\mathbf{x}) - \mathbb{E}_S f_S(\mathbf{x}))^2 \ , \tag{5}$$

Cuando f_S se implementa usando un ensamblado de la forma $f_S(\mathbf{x}) = \frac{1}{N}\sum_{j=1}^{N} f_S^{(j)}(\mathbf{x})$, tenemos que

$$\text{Bias}(f_S(\mathbf{x})) = \frac{1}{N}\sum_j \text{Bias}(f_S^{(j)}(\mathbf{x})) \tag{6}$$

$$\text{Var}(f_S(\mathbf{x})) = \frac{1}{N^2}\sum_j \text{Var}\left(f_S^{(j)}(\mathbf{x})\right) - \frac{1}{N^2}\sum_{j\neq k} \text{Covar}\left(f_S^{(j)}(\mathbf{x}), f_S^{(k)}(\mathbf{x})\right) \ , \tag{7}$$

Si las hipótesis individuales $f_S^{(j)}(\mathbf{x})$ se obtienen del mismo espacio de hipótesis, pero se eligen de modo de estar lo más de-correlacionadas posibles (por ejemplo, son aproximadamente independientes), tenemos

$$\text{Bias}(f_S(\mathbf{x})) \approx \text{Bias}(f_S^{(j)}(\mathbf{x})) \tag{8}$$

$$\text{Var}(f_S(\mathbf{x})) \approx \frac{\text{Var}\left(f_S^{(j)}(\mathbf{x})\right)}{N} \ , \tag{9}$$

es decir, el error de predicción del ensamblado se reduce a través de la estabilización (reducción de varianza) de la predicción.

2. (10 puntos) Demuestre que el Kernel gausiano

$$K(x,z) = e^{\frac{-||x-z||^2}{\sigma^2}},$$

donde $\sigma^2 > 0$ es una constante positiva, es un kernel válido. [Ayuda: Considere que $||x-z||^2 = ||x||^2 - 2x^T z + ||z||^2$.]

Respuesta: Vamos a demostrar primero que si $K(x,z)$ es un kernel válido, entonces $\exp(K(x,z))$ también lo es. En efecto, usando una expansión de Taylor tenemos que:

$$\exp(K(x,z)) = \lim_{i\to\infty} \sum_{j=0}^{i} \frac{1}{j!} K(x,j)^j = \lim_{i\to\infty} K_i(x,z),$$

donde $K_i(x,z) = \sum_{j=0}^{i} \frac{1}{j!} K(x,z)^j$. Notemos que K_i es un kernel válido ya que es una combinación lineal de kerneles validos con escalares positivos. Si denotamos por $\boldsymbol{K}_i$ a la matriz del kernel $K_i(x,z)$, tenemos por lo tanto que $\mathbf{z}^T \boldsymbol{K}_i \mathbf{z} \geq 0$, $\forall i$. Como sabemos que una secuencia de números no negativos $a_i \geq 0$ tiene un límite no negativo $a = \lim_{i\to\infty} a_i \geq 0$, si denotamos por $\exp(\boldsymbol{K})$ la matriz correspondiente al límite $\lim_{i\to\infty} \boldsymbol{K}_i$, tenemos que

$$\mathbf{z}^T \left(\lim_{i\to\infty} \boldsymbol{K}_i \right) \mathbf{z} = \lim_{i\to\infty} \mathbf{z}^T \boldsymbol{K}_i \mathbf{z} \geq 0 \ ,$$

es decir, $\exp(K(x,z))$ es un kernel válido. Para concluir la demostración, basta escribir

$$e^{\frac{-||x-z||^2}{\sigma^2}} = e^{\frac{-||x||^2}{\sigma^2}} e^{\frac{-||z||^2}{\sigma^2}} e^{\frac{2}{\sigma^2} x^T z}$$

Tenemos entonces que $e^{\frac{-||x-z||^2}{\sigma^2}} = f(x) f(z) K(x,z)$ donde los dos primeros términos son escalares positivos y el tercero es un kernel válido, por lo tanto, el kernel gaussiano es un kernel válido.

3. (20 puntos) Derive las ecuaciones del backward pass correspondientes a una red neuronal feed-forward de tres capas, entrenada con la función de pérdida cross-entropy y con función de activación softmax en la capa de salida (asuma que se minimiza el error de entrenamiento).

Respuesta: Denotemos la salida de la red neuronal como $f(\mathbf{x})$. Si la red tiene tres capas, la secuencia de transformaciones ejecutadas por el modelo (forward pass) se resumen en las siguienes ecuaciones:

$$\begin{aligned} f(\mathbf{x}^{(\ell)}) &= g\left(\mathbf{z}^{(\ell)} + \boldsymbol{b}\right) \\ \mathbf{z}^{(\ell)} &= h\left(\mathbf{x}^{(\ell)} + \boldsymbol{c}\right) \ , \end{aligned} \tag{10}$$

donde h es la función de activación de la capa oculta y g denota la transformación softmax

$$g_k(\boldsymbol{\xi}) = \frac{\exp(\xi_k)}{\sum_{k'} \exp(\xi_{k'})} \, . \tag{11}$$

Notemos que

$$g_k'(\boldsymbol{\xi}) = g_k(\boldsymbol{\xi})(1 - g_k(\boldsymbol{\xi})). \tag{12}$$

Ahora, como la red se entrena para minimizar

$$J(\Theta) = -\sum_{\ell} \sum_{k} y_k^{(\ell)} \log f_k(\mathbf{x}^{(\ell)}) \, , \tag{13}$$

las ecuaciones correspondientes al backward pass quedan como sigue. Para la capa de salida,

$$\frac{\partial J(\Theta)}{\partial W_{kj}} = \sum_{\ell} \frac{\partial J(\Theta)}{\partial f_k(\mathbf{x}^{(\ell)})} \frac{\partial f_k(\mathbf{x}^{(\ell)})}{\partial W_{kj}}\,, \tag{14}$$

con

$$\frac{\partial J(\Theta)}{\partial f_k(\mathbf{x}^{(\ell)})} = -\frac{y_k^{(\ell)}}{f_k(\mathbf{x}^{(\ell)})}, \quad \frac{\partial f_k(\mathbf{x}^{(\ell)})}{\partial W_{kj}} = \mathbf{z}_j^{(\ell)} g_k'(\cdot)\,, \tag{15}$$

Para la capa oculta, tenemos

$$\frac{\partial J(\Theta)}{\partial V_{ji}} = \sum_{\ell} \frac{\partial J(\Theta)}{\partial \mathbf{z}_j^{(\ell)}} \frac{\partial \mathbf{z}_j^{(\ell)}}{\partial V_{ji}}\,, \tag{16}$$

El primer término del lado derecho se obtiene usando la regla de la cadena del cálculo multivariado,

$$\frac{\partial J(\Theta)}{\partial \mathbf{z}_j^{(\ell)}} = \sum_{k} \left(\frac{\partial J(\Theta)}{\partial f_k(\mathbf{x}^{(\ell)})} \frac{\partial f_k(\mathbf{x}^{(\ell)})}{\partial \mathbf{z}_j^{(\ell)}} \right)\,, \tag{17}$$

El término $\partial J(\Theta)/\partial f_k$ se encuentra ya disponible de los cálculos correspondientes a la capa sucesiva. Los demás términos se obtienen fácilmente de las ecuaciones del forward pass,

$$\frac{\partial f_k(\mathbf{x}^{(\ell)})}{\partial \mathbf{z}_j^{(\ell)}} = W_{kj} g_k'(\cdot), \quad \frac{\partial \mathbf{z}_j^{(\ell)}}{\partial V_{ji}} = \mathbf{x}_i^{(\ell)} h_j'(\cdot)\,. \tag{18}$$