Universidad Técnica Federico Santa María
Departamento de Informática

INF-578 Máquinas de Aprendizaje

Cuestionario Control 1 II 2017

1. ¿Qué podría decir de un algoritmo de aprendizaje que se diseña para minimizar el error sobre conjunto finito de datos de entrenamiento? Comente Brevemente.

2. Discuta al menos 1 ventaja y 1 desventaja de regularizar el aprendizaje de un modelo lineal usando la norma ℓ_2 (e.g. "ridge regression") versus la norma ℓ_1 (e.g. "lasso").

3. Si tenemos un algoritmo de optimización iterativo que tarda el doble que otro en realizar una iteración. ¿Es cierto que el primer algoritmo convergerá más lento que el segundo?

4. ¿Es cierto o es falso que un algoritmo que escoge una hipótesis en un espacio de hipótesis "más grande" tendrá mejor capacidad de generalización? Por simplicidad, piense en espacios de hipótesis finitos, entendiendo que una colección de funciones "más grande" es una de mayor cardinalidad.

5. Si el z-score asociado a un atributo de entrada es muy cercano a cero. ¿Podría usted afirmar que no existe relación alguna entre dicha variable y la variable de salida?

6. Considere un punto bien clasificado, pero lejos de la frontera de decisión. ¿Por qué la frontera de decisión de la SVM lineal no se ve afectada por este punto, en cambio, la frontera de decisión de la regresión logística si se ve afectada?

7. ([1] 3.4) Considere un modelo lineal de regresión de la forma $f(\mathbf{x}) = \beta_0 + \boldsymbol{\beta}^T\mathbf{x}$, y la suma de errores cuadráticos
$$Q(\boldsymbol{\beta}) = \frac{1}{2}\sum_{\ell=1}^{n}(f(\mathbf{x}^{(\ell)}) - y^{(\ell)})^2.$$
Suponga que una v.a. ϵ_i con distribución $\mathcal{N}(0,\sigma^2)$ se suma independientemente a cada atributo de cada dato $x_i^{(\ell)}$ de entrenamiento. Demuestre que minimizar el valor esperado de $Q(\boldsymbol{\beta})$ sobre el nuevo dataset es equivalente a entrenar el modelo con los datos originales usando *Ridge Regression*, es decir, a minimizar
$$Q'(\boldsymbol{\beta}) = Q(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|^2$$
para algún valor de λ.

8. Suponga que deseamos ajustar un modelo lineal a partir de un conjunto de datos $S = \{(\mathbf{x}^{(\ell)}, y^{(\ell)})\}_{\ell=1}^{n}$ previamente centrados, de modo que podemos omitir el intercepto. Suponga además que la matriz $\boldsymbol{X}$ obtenida al escribir el problema en notación matricial tiene columnas ortogonales.
   (a) Demuestre que, si dos atributos obtenían el mismo "peso" ($\hat{\beta}_i$) en la solución de mínimos cuadrados ordinarios, *Ridge Regression* da mayor peso en el modelo a aquel con mayor z-score.
   (b) Demuestre que, si dos atributos obtenían el mismo "peso" ($\hat{\beta}_i$) en la solución de mínimos cuadrados ordinarios, *Lasso* da mayor peso en el modelo a aquel con mayor z-score.
   (c) Explique en qué difieren los mecanismos utilizados por *Ridge Regression* y *Lasso* para eliminar del modelo aquellos atributos con bajo z-score y en que difieren ambos mecanismos de un método de filtro clásico.

9. ([2] 4.2) Consideremos un problema de clasificación binaria con n datos $\mathbf{x}^{(\ell)} \in \mathbb{R}^d$, donde se tienen n_1 datos de la clase 1 y n_2 de la clase 2.

(a) Muestre que la regla LDA clasifica la clase 2 si

$$x^T\hat{\Sigma}^{-1}(\hat{\mu}_2 - \hat{\mu}_1) > \frac{1}{2}(\hat{\mu}_2 + \hat{\mu}_1)^T\hat{\Sigma}^{-1}(\hat{\mu}_2 - \hat{\mu}_1) - \log(n_2/n_1),$$

y la clase 1 en caso contrario.

(b) Suponga ahora que para cada dato $\mathbf{x}^{(\ell)}$ de clase 1 se define $y^{(\ell)} = -n/n_1$ y para cada dato de clase 2 se define $y^{(\ell)} = -n/n_2$ (es decir, las clases se codifican como $-n/n_1$ y $-n/n_2$ respectivamente). Considere luego la minimización de cuadrados:

$$\sum_{\ell=1}^{n}(y^{(\ell)} - \beta_0 - \beta^T\mathbf{x}^{(\ell)})^2.$$

Muestre que después de simplificaciones, la solución β satisface

$$\left[(n-2)\hat{\Sigma} + n\hat{\Sigma}_B\right]\beta = n(\hat{\mu_2} + \hat{\mu_1}),$$

donde $\hat{\Sigma}_B = \frac{n_1 n_2}{n^2}(\hat{\mu_2} + \hat{\mu_1})(\hat{\mu_2} + \hat{\mu_1})^T$

(c) Continuando con el ejercicio anterior, muestre que $\hat{\Sigma}_B\beta$ está en la dirección $(\hat{\mu_2} - \hat{\mu_1})$, es decir

$$\hat{\beta} \propto \hat{\Sigma}^{-1}(\hat{\mu_2} - \hat{\mu_1})$$

Por lo tanto, el coeficiente de la regresión de de mínimos cuadrados es idéntico al coeficiente de LDA escalado por una constante.

References

[1] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[2] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2009.