Universidad Técnica Federico Santa María
Departamento de Informática

INF-578 Máquinas de Aprendizaje

Cuestionario Control 1 II 2017

Los ejercicios marcados con (★) son solo para alumnos de postgrado.

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 el rol del conjunto de entrenamiento (*training set*), conjunto de validación (*validation set*) y conjunto de pruebas (*test set*) en aprendizaje automático.

3. Explique en qué consiste la técnica de "validación cruzada" (k-fold cross validation) en aprendizaje automático. Escriba un mini-programa (pseudo-código) que implemente esta técnica para la selección de un parámetro estructural (hiper-parámetro) genérico λ.

4. Considere el modelo estándar de regresión lineal $f(\mathbf{x}) = \beta^T\mathbf{x} + b$ en $\mathbb{R}^d$, con $d = 1000000$. Explique porqué, pese a su simplicidad, el aprendizaje de este modelo desde un conjunto de ejemplos $S_E = \{(\mathbf{x}^{(\ell)}, y^{(\ell)})\}_{\ell=1}^{100}$ es suceptible de sobreajuste (*overfitting*). ¿Cómo lo podría prevenir?

5. Explique qué se entiende por datos "dispersos" (*sparse data*) señalando la importancia de explotar explícitamente esta característica en un problema de aprendizaje automático. Indique además porqué la normalización de datos resulta particularmente complicada en este caso.

6. Explique la diferencia entre *selección de atributos* y *reducción de dimensionalidad*. Refiérase a una ventaja y a una desventaja de cada enfoque.

7. Explique la diferencia entre un enfoque de tipo *filtro* y uno tipo *wrapper* para selección de atributos. Escriba el pseudo-código correspondiente a un método de cada categoría aplicado a un problema de regresión y determine su complejidad computacional.

8. 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").

9. Explique brevemente el supuesto denominado "Bayesiano Ingenuo" (*Naive Bayes*) en el diseño de clasificadores probabilistas.

10. Explique brevemente las elecciones de diseño que distinguen LDA (*Linear Discriminant Analysis*) de QDA (*Quadratic Discriminant Analysis*), demostrando que las primeras derivan en fronteras de clasificación lineales, mientras que las segundas derivan en en fronteras de clasificación cuadráticas. Discuta también las ventajas y desventajas de la diferencia obtenida.

11. ¿Es cierto o es falso que si resolvemos el problema de regresión lineal mediante mínimos cuadrados ordinarios con gradiente descendente podemos encontrar varios óptimos locales? Justifique seriamente.

12. 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?

13. (★) ¿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.

14. (★) 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?

15. (★) 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?

16. Calcule la esperanza y la matriz de covarianza del estimador $\hat{\beta}$ optenido usando mínimos cuadrados.

17. ([2] 2.9) Considere un modelo de regresión lineal con p parámetros, obtenido usando el método de mínimos cuadrados a partir de una muestra aleatoria

$$S_E = \{\mathbf{x}^{(1)}, y^{(1)}), \ldots, (\mathbf{x}^{(n)}, y^{(n)})\},$$

y denote por $\hat{\beta}$ el estimador de mínimos cuadrados. Suponga que tenemos el conjunto de test $S_T = \{(\tilde{\mathbf{x}}^{(1)}, \tilde{y}^{(1)}), \ldots, (\tilde{\mathbf{x}}^{(m)}, \tilde{y}^{(m)})\}$ obtenido aleatoriamente desde la misma población. Si

$$R_{tr}(\beta) = \frac{1}{n}\sum_{\ell=1}^{n}(y^{(\ell)} - \beta^T\mathbf{x}^{(\ell)})^2$$

es el error cuadrático sobre el conjunto de entrenamiento y

$$R_{ts}(\beta) = \frac{1}{m}\sum_{\ell=1}^{m}(\tilde{y}^{(\ell)} - \beta^T\tilde{\mathbf{x}}^{(\ell)})^2$$

es el error cuadrático sobre el conjunto de prueba. Demuestre que:

$$E[R_{tr}(\hat{\beta})] \leq E[R_{ts}(\hat{\beta})]\,.$$

18. ([2] 4.5) Considere un problema de clasificación con datos $\{(\mathbf{x}^{(\ell)}, y^{(\ell)})\}_{\ell=1}^{n}$ tal que $\mathbf{x}^{(\ell)} \in \mathbb{R}^1$ e $y^{(\ell)} \in \{0, 1\}$. ¿Cuáles serían las ecuaciones correspondientes a un clasificador logístico que se desea entrenar para abordar este problema? ¿Cuántos parámetros libres existen en este caso? Suponga ahora, que existe un punto $\mathbf{x}_0 \in \mathbb{R}$ tal que $\forall \mathbf{x}$, $P(y = 1|\mathbf{x}) = 1$ si $\mathbf{x} < \mathbf{x}_0$, y $P(y = 1|\mathbf{x}) = 0$ si $\mathbf{x} > \mathbf{x}_0$. Demuestre que en este caso, la función de verosimilitud asociada al modelo no es acotada y que por lo tanto el estimador máximo verosímil es degenerado.

19. (★) ([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.

20. (★) 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.

21. (★) ([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.