Taller II

Tópicos en Aprendizaje de Maquinas

2021 I

Alvaro J. Riascos Villegas

Jose Sebastian Ñungo

**Instrucciones:** Para algunos de los ejercicios es posible que requieran investigar un poco. Se recomienda usar el libro James, G. Witten, D. Hastie, T. Tibshirani, R (2013). An Introduction to Statistical Learning with Applications in R. Springer New York Inc. Available in: http://www-bcf.usc.edu/gareth/ISL/3

1. *Métodos de Regularización/Penalización*: Para cada inciso de este punto se requiere justificar cada respuesta.
2. Después de aplicar un parámetro de penalización en una regresión lineal, se encuentra que algunos coeficientes se han vuelto cero. ¿Cuál de los siguientes tipos de penalización se ha usado?

1) Norma L0

2) Norma L1

3) Norma L2

4) Norma L0 o L1

5) Ninguna de las anteriores

1. ¿Cuál de los siguientes métodos de regularización son invariantes a la escala?

1) L0 y L2 pero no L0

2) L1 y L2 pero no L0

3) L0 pero no L1 o L2

4) L2 pero no L0 o L1

5) Ninguna de las anteriores

1. Suponga que ha construido un modelo de regresión complejo e intenta reducir su complejidad usado la regresión Lasso. De los siguientes enunciados escoja el correcto:

1) Cuando el *λ* es muy grande, el sesgo es bajo y la varianza es baja.

2) Cuando el *λ* es muy grande, el sesgo es bajo y la varianza es alta.

3) Cuando el *λ* es muy grande, el sesgo es alto y la varianza es baja.

4) Cuando el *λ* es muy grande, el sesgo es alto y la varianza es alta.

1. ¿En qué situaciones la regresión ridge se prefiere a la regresión lasso? Establezca ventajas y desventajas de los dos métodos.
2. Este ejercicio está basado en la lectura *Three Learning Principles* que se encuentra en la pagina web del curso.
3. ¿Cuál es el principio de la navaja de Occam y de qué forma se aplica al proceso de minería de datos?
4. ¿Cuál es el problema fundamental cuando se enfrenta a una alta dimensionalidad en los datos y de qué forma los algoritmos vistos en clase se encuentran en la misma línea filosófica del principio de Occam como parte de la solución? Explique y elabore ejemplos.
5. ¿A qué se refiere el sesgo de muestreo en la minería de datos? ¿Por qué es importante? ¿Qué tipos de medidas pueden tomarse para no producir este tipo de sesgo en los ejercicios de predicción?
6. Explique la relación entre el sesgo de muestreo y *data snooping*. De acuerdo a lo visto en clase, explique formas de evitar caer en la trampa de *data snooping*.
7. *Regresión con Métodos de Selección y Regularización.* Para este ejercicio se utilizará la base de datos *cancer\_mama* perteneciente a la Universidad de Wisconsin. Los datos contienen el diagnóstico de cáncer de mamá en 569 mujeres. Las características del diagnóstico corresponden a características obtenidas mediante un análisis de imágenes digitalizadas de masas en los senos. Para mayor información pueden consultar: [UCI Machine Learning Repository: Breast Cancer Wisconsin (Original) Data Set](https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29). El objetivo de este ejercicio es predecir el diagnóstico de cáncer de mamá (*Diagnosis*) con base en las características de la masa observada. Este ejercicio debe ser contestado en un cuaderno de Jupyter siendo muy claro en donde está respondiendo cada uno de los siguientes ítems y con su código asociado. Recuerde que puede utilizar todos los paquetes de Python vistos en clase.
8. Si es necesario, convierta los datos a los formatos correctos. Obtenga una descripción de la base de datos. Describa los resultados.
9. Implemente correctamente un modelo de regresión logística como un problema de predicción, usando como variable a predecir *Diagnosis*. ¿Qué piensa de los resultados?
10. Ajuste una regresión logística con selección de variables usando métodos de selección de variables *Forward, Backward y Stepwise* para encontrar el mejor modelo. Para cada uno de los métodos calcule la matriz de confusión, *accuracy*, *precision*, *recall*. ¿Qué diferencia observa entre las métricas de desempeño de los diferentes modelos? Explique.
11. Presente en una tabla las variables de la regresión del punto (b) y las variables finales de los tres métodos de selección.
12. Implemente las técnicas de regularización Ridge y Lasso bajo un problema de clasificación y determine cuál tiene mejor desempeño según las métricas de desempeño descritas en el inciso (c). ¿el modelo elegido contiene todas las variables de la base de datos? ¿Por qué?