Aprendizaje Estadístico 2014. Examen 1

1. Sea Y una variable aleatoria independiente de $X = (X_1, \ldots, X_p)$. Calcula un predictor de Y, en función de X, que de la mínima pérdida cuadrática para predecir la salida Y dado X.

2. Sea $X = (X_1, \ldots, X_p)$, donde X es uniforme en la bola $\{x \in R^p | ||x||^2 < 1\}$, y $\mathcal{L}$ es una muestra aleatoria de tamaño N de la distribución de X. Considera la variable aleatoria $D = \text{mín}_{x \in \mathcal{L}} ||x||$, que da la mínima distancia al origen entre los puntos de la muestra $\mathcal{L}$. Calcula la mediana de D y explica qué tiene qué ver esto con la maldición de las dimensiones.

3. Sea $x_i = i/20$ para $i = 0, \ldots, 20$ **fijos**, y supón que cada respuesta se construye como $y_i = |x_i - 1/2| + \epsilon_i$ con $\epsilon_i \sim N(0, 1)$ independientes. Usaremos k-vecinos más cercanos para predecir Y.

   Calcula la descomposición del error esperado de predicción para $x_{10} = 0.5$ en irreducible, sesgo y varianza para $k = 1, 3, 6$. Describe el comportamiento de cada componente cuando k toma estos distintos valores.

4. Demuestra que la solución del problema ridge está dada por

$$\hat{\beta}^{ridge} = (\underline{X}^t \underline{X} + \lambda I)^{-1} \underline{X}^t y$$

5. Considera el problema ridge. Demuestra que si las variables están centradas, y no penalizamos la ordenada al origen β_0, entonces $\hat{\beta}_0^{ridge}$ es la media de las y_i, independientemente de la λ seleccionada.

6. Muestra que las estimaciones de regresión ridge pueden obtenerse por mínimos cuadrados en un conjunto de datos aumentado. Aumentamos $\underline{X}$ con p renglones adicionales $\sqrt{\lambda} I$, y aumentamos y con p ceros. Muestra que agregando estos datos artificiales el proceso de ajuste está forzado a encoger los coeficientes hacia cero.

7. Suponemos que la matriz $\underline{X}$ de entradas es ortogonal. Escribe la solución de ridge para $\lambda > 0$ en términos de la solución de mínimos cuadrados. ¿En qué sentido se encogen los coeficientes? Repite para lasso (sugerencia: demuestra que el problema de lasso, en este caso, se puede resolver variable por variable. Divide en casos para λ y resuelve para un coeficiente).

8. Considera el siguiente problema: tenemos $p = 5000$ predictores y una muestra de $N = 50$ casos, donde la respuesta es binaria 0-1. Un analista seleccionó las 100 variables de entrada que tienen mayor correlación con la respuesta. Con estas 100 variables, construyó entonces un predictor ridge (o lasso o k-vmc) donde escogió el parámetro de complejidad por validación cruzada. Su estimación por validación cruzada del error es de 0.03 de casos mal clasificados. Después de un tiempo, alguien le dijo que de hecho la respuesta se construyó con volados de una moneda justa. Discute: ¿Cuál va a ser el desempeño futuro del predictor del analista? ¿Es

realista la estimación de validación cruzada del analista? ¿Es una correcta aplicación de validación cruzada? Si no lo es, ¿cómo debería hacerse?

9. (Separabilidad en regresión logística) Supón que tenemos una sola entrada x, y que los casos de tipo 1 cumplen $x > 0$ y los casos tipo 0 cumplen $x < 0$.

   - Explica por qué el problema de minimización de la devianza no tiene solución. ¿Qué pasa con los coeficientes de $p_1(x;\beta) = h(\beta_0 + \beta_1 x)$? (una gráfica es suficiente).
   - Muestra que si regularizamos (por ejemplo con ridge), entonces el problema de minimización de la devianza penalizada tiene solución única.

10. Considera un problema de clasificación en 2 clases con matriz de pérdida dada por $L_i i = 0$, $L_{12} = 1$, $L_{21} = a$, donde $a > 1$ es una constante. Supón que la clase 2 es el *positivo*.

    a) Calcula el clasificador que minimiza la pérdida esperada. Escribe este clasificador en términos de un punto de corte para $P(G = 2|X = x)$.

    b) Calcula cotas para la especificidad y sensibilidad del clasificador óptimo (inciso anterior). En base a estas cotas, explica qué pasa con estas dos cantidades cuando a tiende a infinito.

11. (Entregar - regresión) Queremos predecir el salario de un beisbolista en función de varias estadísticas que describen su desempeño (datos *Hitters* del paquete *ISLR*). Construye un modelo de regresión lineal, uno de regresión ridge, uno de regresión lasso y uno de 1-vecino más cercano (estandariza las variables para 1-vecino más cercano). Escoge los modelos de ridge y lasso usando validación cruzada. Aparta una muestra de prueba de al menos cien casos escogidos al azar para hacer una evaluación final de tus 3 modelos. Reporta los errores de entrenamiento y de prueba en cada caso.

12. (Entregar) Consideramos los datos de *SAheart* del paquete *ElemStatLearn*, donde queremos predecir enfermedad del corazón (*chd*) en términos de otras variables como consumo de tabaco, obesidad, edad, etc. Separa una muestra de entrenamiento de 300 casos y deja el resto para prueba. Construye un modelo de regresión logística (usa ridge o lasso), y selecciona con validación cruzada el parámetro de regularización.

    - ¿Cuál es la tasa de clasificación incorrecta de tu modelo? ¿Cómo se compara con la tasa base de clasificación incorrecta?
    - ¿Cuál es la especificidad y sensibilidad de tu modelo?
    - Construye curvas ROC de entrenamiento y prueba para el modelo seleccionado.

13. (Entregar-simulación) Considera el siguiente modelo: $x \sim U(0,1)$, y $y = |x - 1/2| + \epsilon$ con $\epsilon \sim N(0,1)$. Queremos usar 3 vecinos más cercanos para predecir y en función de x.

    Produce 100 muestras de entrenamiento de tamaño $n = 30$, y una muestra de prueba grande. Usando error cuadrático medio, haz una gráfica que compare el error de entrenamiento con el error de predicción (condicional a la muestra de entrenamiento) para cada muestra de entrenamiento. Usa estas 100 muestras para estimar el error de predicción no condicional. ¿Hay mucha variabilidad del error de predicción condicional obtenido alrededor del error no condicional?