Overfit, cross validation y bootstrap

Test de Lectura

Como discutiesemos, estos *tests de lectura* son una serie de preguntas simples que permiten monitorear si estan siguiendo efectivamente las clases. No seran corregidos ni evaluados, los administraran directamente ustedes. Cualquier duda que tengan sera discutida en nuestras sesiones sincronicas, la semana siguienta a publicado el material, luego de que hayan hecho un esfuerzo en trabajarlo, no antes. Es muy importante que trabajan este material y lo discutan entre ustedes.

1. V o F: la ley de grandes numeros dice que el promedio muestral tiene distribucion normal, en el limite.
2. Por que en econometria clasica de minimizar el ECM es el de minimizar la varianza?
3. Intuitivamente, como se 've' el problema de overfit en el grafico del slide de la pagina 8?
4. En el caso del polinomio, por que el sesgo existe cuando $p < p*$?
5. Y por que desparece cuando $p > p*$?
6. Para que usamos la traza en la demostracion del slide 11?
7. V o F: en el caso del polinomio, para cualquier grado del polinomio la varianza tiende a ser nula cuando la cantidad de observaciones tiende a infinito.
8. En base a lo anterior, pensa por que lo del trade/off sesgo-varianza es esencialmente un problema de muestra finita.
9. Si el predictor es sesgado, que pasa con el sesgo a medida que la muestra crece?
10. Por que R^2 no funciona para medir la capacidad de pronostico fuera de la muestra?
11. En el ejemplo del grafico del slide de la pagina 8, como seria un grafico del R^2 en funcion del grado del polinomio ajustado?
12. En cross validation, cual es el problema de usar un K muy pequeño y uno muy grande?

13. Dificil: si para hacer cross validation estimamos el modelo K veces, una idea es mejorar el estimador inicial, reemplazandolo por uno que toma promedio de los K modelos estimados en el proceso de hacer cross validation. Pensa por que esta estrategia en general no funciona.

14. V o F: en cross validation, cuando $K = n$ el modelo es estimado n veces.

15. Explica intuitivamente cual es el procediminto usado para hacer el grafico del slide de la pagina 22.

16. Supongamos que queres calcular el error estandar de la mediana muestral. Explica como lo harias con bootstrap.

17. Supone que tenes datos de ingresos de 1,000 personas y que se considera que una persona es pobre si sus ingresos estan por debajo de 130 y que la proporcion de personas pobres te da 35. Explica detalladamente como construirias una intervalo de confianza al 90 % por bootstrap usando $B = 300$.

18. A proposito no hable de cuantas replicaciones boostrap son necesarias/suficientes. Una parte enorme de la tarea del cientifico de datos es googlear. Googlealo.