Introducción al aprendizaje estadístico

Índice

El aprendizaje estadístico se refiere a una conjunto de herramientas y entendimiento de conjuntos de datos complejos. Se ha desarrollado recientemente en el area de estadística y se mezcla con el desarrollo en pararelo de las ciencias computaciones, en particular, con aprendizaje de máquina.

1. Métodos de Remuestreo

Remuestro o evaluación del método de aprendizaje.

Los métodos de remuestro son una herramienta indispensable en la estadística moderna. Éstos involucran la toma repetida de muestras de un conjunto de entrenamiento y reajustan el modelo de interés en cada muestra para obtener información adicional acerca del modelo ajustado.

Llamamos evaluación del modelo(model assessment) al proceso de evaluar el desempeno del modelo y al proceso de seleccionar el nivel adeacuado de flexibilidad le llamamon selección del modelo (model selection).

En este capítulo veremos dos de los métodos más usados: Validación cruzada y bootstrap.

1.1. Validación cruzada. Recordando que el error de prueba (test error) es el promedio de error que resulta de utilizar un método de aprendizaje estadístico al hacer una predicción de una nueva observación. El error de prueba puede ser calculado si tenemos un conjunto disponible, desafortunadamente no siempre es el caso. En la validación se estima el error de prueba usando una parte de los datos para el entrenamiento y la otra parte restante para la prueba.

En las siguiente subsecciones consideraremos una variable de respuesta cuantitativa. Mas adelante veremos el caso de la respuesta cualitativa y como es de esperarse, los conceptos se mantienen igual sin importar el tipo de respuesta.

*Como hemos visto, el error de entrenamiento tiende a subestimar el error de prueba (test error).

1.1.1. El enfoque del conjunto de validación. El conjunto de datos se divide en dos subconjuntos (comparables en tamano): el de entrenamiento y el de validación. El modelo es ajustado usando el conjunto de entrenamiento y se calcula el MSE en el conjunto de validación, así obtenemos un estimado de la tasa de error de prueba.

Si recordamos el ejemplo del conjunto de datos Auto, vimos que el mpg se puede predecir usando horsepower. También notamos que el modelo cuadrático ($horsepower^2$) era mejor que el lineal. Esto se resolvió usando $p - valores$ pero también se puede resolver usando el método de validación. Si calculamos el MSE en el conjunto de validación del modelo cuadrático vemos que es mucho menor que el MSE en el modelo lineal. Así tenemos una forma diferente de encontrar un mejor modelo.

Este enfoque de validación tiene dos desventajas:

1. La estimación de la tasa de error de prueba puede ser muy variable, dependerá de que observaciones son incluidas en el conjunto de entrenamiento y de cuales son incluidas en el conjunto de validación.

2. Como solo una parte de los datos es usado en el ajuste del modelo (trainning set) y como los métodos de estadística tienden a desempenarse peor cuando son entrenados con pocas observaciones, esto sugiere que la tasa de error en el conjunto de validación tiende a sobreestimar la tasa de error para el modelo en todo el conjunto de datos.

El método de validación cruzada que veremos más adelante, refina este método de validación.

1.1.2. Validación cruzada Deja-uno-fuera. Similar al anterior, si tienemos n datos, los dividimos en dos subconjuntos: uno de $n-1$ datos y el otro de un solo dato. Esto es, $\{(x_2, y_2), ..., (x_n, y_n)\}$ son el conjunto de entrenamiento y $\{(x_1, y_1)\}$ es el conjunto de validación. Así $MSE_1 = (y_1 - \hat{y_1})^2$ es aproximadamente un estimador insesgado del error de prueba, pero que es muy variable pués solo está creado por una observación. El proceso se puede repetir dejando cada una de las observaciones por fuera y calculando el correspondiente MSE_i. Nuestro estimador para el MSE será el promedio esos n errores:

$$CV_{(n)} = \frac{1}{n} \sum_{i=1}^{n} MSE_i$$

Las ventajas de este método sobre el de validación las describimos a continuación. Tiene menos sesgo pués utiliza $n-1$ observaciones, casi todo el conjunto ! Así no sobreestimamos la tasa de error de prueba. En el método de validación los resultados dependían del tipo de observaciones en cada conjunto y por consecuente diferentes. En este método no hay aleatoriedad en la división de los conjuntos, asi obtenemos los mismos resultados.

La validación cruzada Deja-uno-fuera es muy general y puede utilizarse en cualquier modelado de predicción, como ejemplo: Regresión logística, Análisis linea discriminante o algún otro que vemos más adelante.

1.1.3. Validación cruzada K-fold. Este método generaliza el anterior. Dado el conjunto de datos, lo dividimos en k subgrupos. El primer grupo será el de validación y utilizamos los demás $k-1$ grupos para ajustar el modelo. Encontramos MSE_1 con el grupo que quedó fuera. Este procedimiento lo repetimos k veces y encontramos $MSE_2, MSE_3, ..., MSE_k$. El estimador lo encontramos usando estos valores:

$$CV_{(k)} = \frac{1}{k} \sum_{i=1}^{k} MSE_i$$

El método anterior (Deja-uno-fuera) es es un caso particular cuando $k = 1$. Lo más común es usar $k = 5$ o $k = 10$, pues es computacionalmente accesible y ademas de algunas ventajas en el sesgo-varianza.

En general utilizamos la validación cruzada para saber que tan bien se desempena un método estístico, esto es, nos interesa conocer el test MSE (evaluación del modelo). En otras ocasiones, estaremos interesados en el punto mínimo de la curva del estimador del test MSE para conocer cual es el mejor método a utilizar (selección del modelo).

El k-fold tiene ventajas sobre el Deja-uno-fuera, ademas de la computacional tiende a dar estimados de la tasa de error de prueba mas precisos. Aunque el segundo metodo tiene menor sesgo por usar los $n-1$, su varianza es mas alta. Tipicamente utilizaremos la validación cruzada con $k = 5$ o $k = 10$ cuando estos valores muestren empíricamente que el error de prueba estimado no sufre de mucho sesgo ni mucha varianza.

1.1.4. Validación cruzada en clasificación. En los problemas de regresión utilizamos el MSE para cuantificar los erroes. Ahora, utilizaremos el número de clasificaciones hechas incorrectamente. En el método de Deja-uno-fuera, el la tasa de error

toma la formad de

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} Err_i$$

donde $Err_i = I(y_i \neq \hat{y}_i)$. Para el método k-fold, la tasa de error se define de manera análoga.

1.2. Bootstrap. Esta es una herramienta estadística ampliamente aplicable y extremadamente poderosa que puede ser usada para cuantificar la incertidumbre asociada a una estimador o un método de aprendizaje. Como ejemplo, puede ser utilizado para estimar los errores estándares de los coeficientes de la regresión lineal. Aunque no es muy útil pués los paquetes de software los pueden generar. La ventaja de bootstrap es que puede medir la variabilidad de los métodos de aprendizaje estadístico que son difíciles de obtener o que los paquetes de software no los generan automáticamente.

En esta sección se illustra el bootstrap un ejemplo.

Tarea para el Lunes 27 de Abril.

1. Explicar el ejemplo Boostrap de la pagina 187.

2. Hacer el Lab 5.3 de la pagina 190. Correr el código en R y explicar con sus palabras (dentro del codigo R) de que se trata la practica.