Evaluación de la Performance en Modelos de Clasificación

EST 266 - Minería de Datos

Pontificia Universidad Católica del Perú

Resumen I

Desempeño de un clasificador I

- Un clasificador necesita datos (datos de entrenamiento) para construir el modelo:
  - En la regresión logística: estimar los coeficientes del modelo.
  - En árboles de clasificación: determinar los nodos y umbrales.
  - En máquinas de soporte vectorial: encontrar los vectores de soporte.
- Para evaluar el clasificador construido sobre la base de entrenamiento, otro conjunto de datos (datos de prueba) es necesario.
- El clasificador etiqueta cada objeto (instancia) en la base de datos y esta clasificación puede ser evaluada usando diversas medidas de evaluación.

Tasa de Error de Mala Clasificación I

- Dado un clasificador d, su tasa de error de mala clasificación $R(d)$ es la probabilidad de que d clasifique incorrectamente un objeto de una muestra (muestra de prueba) obtenida posteriormente a la muestra de entrenamiento.
- También es llamado el error verdadero.
- Es un valor desconocido que necesita ser estimado.

Error por resustitución o error aparente I

- Smith (1947).
- La proporción de observaciones incorrectamente clasificadas por el método utilizando los datos de la muestra de entrenamiento.
- Estimador demasiado optimista que subestima el verdadero error.
- Puede conducir a conclusiones erróneas cuando el número de instancias no es demasiado grande comparado con el número de atributos.
- Suele tener un sesgo grande.

Método de Retención I

- Usada por la librería `Rattle`.
- Se extrae un porcentaje de los datos (usualmente $70\,\%$) que es considerada como **muestra de entrenamiento**.
- Los datos restantes son considerados como muestra de prueba.
- El modelo es estimado en la muestra de entrenamiento y el error es calculado en la muestra de prueba.
- Una estrategia alternativa es tomar el $30\,\%$ de las observaciones restantes y dividirlas en dos grupos iguales: $15\,\%$ para una **muestra de validación** y el $15\,\%$ para la **muestra de prueba**.
- El conjunto de validación es usado para probar diferentes configuraciones de los parámetros del modelo o diferentes alternativas de variables mientras se está construyendo el modelo. Proporciona una estimación de que tan bien el modelo se desempeñará con nuevas observaciones.

Método de Retención II

- Es importante resaltar que este conjunto de datos no es usado para construir el modelo ni para estimar el error o evaluarlo y compararlo con otros modelos.
- El conjunto de datos de prueba solamente será usado para predecir el error insesgado con los resultados finales.
- Desventajas:
  - Las instancias incluidas en la muestra de prueba pueden ser muy fáciles o muy difíciles de clasificar (subestimar o sobrestimar el verdadero error).
  - Instancias que son esenciales para construir el modelo pueden no estar incluidas en el conjunto de entrenamiento.
- Para mitigar este efecto, el experimento puede ser repetido varias veces y luego se promedia la tasa de error en las muestras de prueba. Sin embargo, existe la posibilidad de que ambos problemas se presenten.

Estimación dejando uno afuera (LOO) I

- Lachenbruch (1965).
- En este caso una observación es omitida de la muestra de entrenamiento.
- Luego, el clasificador es construido y se obtiene la predicción para la instancia omitida.
- Se registra si la instancia fue correcta o incorrectamente clasificada.
- Se debe repetir el proceso con todas las instancias.
- Finalmente se estima el error calculando la proporción de instancias mal clasificadas.
- Este estimador tiene poco sesgo pero su varianza tiende a ser grande.

Validación cruzada (K-CV) I

- Stone (1974).
- La muestra de entrenamiento es dividida en K partes.
- El clasificador es construido utilizando todas las partes menos una $(K-1)$.
- La parte omitida es considerada como la muestra de prueba y se hallan las predicciones de cada una de sus instancias.
- La tasa de error de mala clasificación CV es hallada sumando el número de malas clasificaciones en cada parte y dividiendo el total por el número de instancias en la muestra de entrenamiento (promedio de las K partes).
- La principal ventaja de esta técnica es que cada instancia es evaluada exactamente una vez.
- Este estimador tiene poco sesgo pero una varianza alta.

Validación cruzada (K-CV) II

- La elección de K depende del balance entre sesgo-variabilidad del clasificador (Kohavi, 1995):
  - Para valores pequeño de K, el número de muestras de entrenamiento será menor, y las clasificaciones tendrán más sesgo dependiendo de que tanto cambia el rendimiento del clasificador con las instancias incluidas en el conjunto de entrenamiento y del tamaño de la muestra.
  - Para valores grandes de K, el procedimiento se mueve más variable debido a la fuerte dependencia sobre la muestra de entrenamiento.
  - Usualmente se utiliza K entre 5 y 10 (10-fold-cross-validation).
- Para reducir la variabilidad se repite la estimación varias veces.

Bootstrapping I

- Efron (1983).
- Se generan muestras con reemplazo con las cuales se construye el clasificador.
- La idea es reducir el sesgo del error aparente.
- Este estimador es casi insesgado pero tiene una varianza alta.
- Su costo computacional es alto.

Vehicle

A data frame with 846 observations on the following 19 variables (library dprep).

- V1 Compactness
- V2 Circularity
- V3 Distance Circularity
- V4 Radius ratio
- V5 pr.axis aspect ratio
- V6 max.length aspect ratio
- V7 scatter ratio
- V8 elongatedness
- V9 pr.axis rectangularity
- V10 max.length rectangularity
- V11 scaled variance along major axis
- V12 scaled variance along minor axis
- V13 scaled radius of gyration
- V14 skewness about major axis
- V15 skewness about minor axis
- V16 kurtosis about minor axis
- V17 kurtosis about major axis
- V18 hollows ratio
- V19 Type of vehicle: a double decker bus, Cheverolet van, Saab 9000 and an Opel Manta 400.

Ejemplo

```
#LDA para el conjunto Vehicle que tiene 4 clases

#LDA para el conjunto Vehicle que tiene 4 clases
library(MASS)
load("vehicle.rda")

# Error por resustitucion

ldaveh=lda(vehicle[,1:18],vehicle[,19])
predict(ldaveh)$posterior
predict(ldaveh)$class
#Matriz de confusion
table(vehicle[,19],predict(ldaveh)$class)
mean(vehicle[,19]!=predict(ldaveh)$class)
```

Ejemplo

```
#Por el metodo de retencion
set.seed(666)
n<-dim(vehicle)[1]
ntrain <- round(0.7*dim(vehicle)[1],0)
train = sample(n,ntrain)
ldaveh.ret=lda(vehicle[train,1:18],vehicle[train,19])
y<-vehicle[-train,19]
yhat<-predict(ldaveh.ret,
                    newdata=vehicle[-train,1:18] )$class
table(y,yhat)
mean(y!=yhat)
```

Ejemplo

```
#Por el metodo, dejando uno afuera
ldaveh.cv=lda(vehicle[,1:18],vehicle[,19],CV=TRUE)
y<-vehicle[,19]
yhat<-ldaveh.cv$class
table(y,yhat)
mean(y!=yhat)
mean(vehicle[,19]!=ldaveh.cv$class)
```

Evaluación de las Clases Predichas I

Predichos	Observados	
	Eventos	No Eventos
Eventos	TP	FP
No Eventos	FN	TN

Figura: Matriz de confusión para un problema de clasificación con dos clases (eventos y no eventos). Las celdas de la tabla indican el número de verdaderos positivos (TP), falsos positivos (FP), verdaderos negativos (TN) y falsos negativos (FN)

Evaluación de las Clases Predichas II

- En la matriz de clasificación mostrada anteriormente la fila superior de la tabla corresponde a las observaciones predichas como eventos.
- Algunas son predichas correctamente (verdaderos positivos o TP) mientras otros de manera inadecuada (falso positivos o FP).
- De manera similar la segunda fila contiene las predicciones negativas como verdaderos negativos (TN) y los falsos negativos (FN).
- El conjunto de datos de evaluación consta de P observaciones Positivas y N Negativas tal que:

$$TP + FN = P$$

$$TN + FP = N$$

Evaluación de las Clases Predichas III

- La métrica más simple es el ratio de la exactitud total.

$$acc(d) = \frac{TP + TN}{P + N}$$

- o, siendo pesimistas, el ratio de error: $1 - acc(d)$
- Este patrón es un indicador de que el modelo tiene una pobre calibración y también desempeño.
- Existen ciertas desventajas al usar estas estadísticas:
  - En primer lugar, la exactitud no realiza distinciones entre el tipo de error que fue cometido. Ejemplo: En filtros de spam, el costo de borrar erróneamente un e-mail importante es mucho mayor que los costos de permitir que un emailspam pase el filtro. Una discusión sobre este problema puede ser revisado en Provost et al. (1998).
  - En segundo lugar, es importante considerar la frecuencia total de cada clase.

Evaluación de las Clases Predichas IV

- ¿Cuál es el ratio de exactitud que debe ser utilizado como *benchmark* para determinar si un modelo se desempeña adecuadamente?. Se puede usar un ratio no informativo.
- Un ratio no informativo es el ratio de exactitud que se podría alcanzar sin usar un modelo.
- Existen varias formas para definir este ratio:
  - Para un conjunto de datos con C clases, la definición más simple, basada solamente en el azar, es $1/C$. Esta definición no toma en cuenta las frecuencias relativas de las clases en el conjunto de entrenamiento.
  - Una definición alternativa es el porcentaje de la clase de mayor frecuencia en el conjunto de entrenamiento.Modelos con una precisión mayor a este ratio pueden considerarse como razonables.
  - Sobre el efecto severo de clases no balanceadas y posibles medidas remediales ver Kuhn y Johnson (2013).

Evaluación de las Clases Predichas V

- El coeficiente Kappa fue originalmente diseñado para evaluar la concordancia entre dos evaluadores (Cohen, 1960).
- Kappa toma en cuenta la precisión que sería generada por causas aleatorias. Se encuentra definido como:

$$Kappa = \frac{O - E}{1 - E}$$

  donde O es la precisión observada y E la precisión esperada basada sobre los totales marginales de la matriz de confusión.
- El estadístico puede tomar valores entre -1 y 1. Un valor de 0 indica que no hay concordancia entre las clases observadas y las pronosticadas, mientras un valor de 1 indica un perfecta concordancia de la predicción del modelo con las clases observadas.
- Valores negativos indican que las predicciones están en una dirección opuesta, aunque esto raramente ocurre.

Evaluación de las Clases Predichas VI

- Dependiendo del contexto, valores de Kappa entre 0.30 a 0.50 indican una razonable concordancia. Suponga que la precisión para un modelo es alta (90 %) pero la precisión esperada es también alta (85 %), el estadístico Kappa podría mostrar una concordancia moderada (Kappa=1/3) entre las clases observadas y pronosticadas.
- El estadístico Kappa puede ser también extendido para evaluar la concordancia en problemas con más de dos clases. Cuando hay un ordenamiento natural de las clases (p. ej. bajo, medio y alto), una forma alternativa es el estadístico Kappa ponderado puede ser usado para representar sanciones más severas en errores que se alejen más del verdadero resultado. Revisar Agresti (2002) para mayores detalles.

Clasificación Binaria I

- Para dos grupos existen estadísticas adicionales que pueden ser relevantes cuando una clase es interpretada como un evento de interés.
- La sensibilidad o recuperación de un modelo es el ratio en que el evento de interés es predicho correctamente para todas las muestras que contienen el evento.

$$sensibilidad(d) = \frac{TP}{TP + FN}$$

- La sensibilidad es usualmente considerada como el ratio de verdaderos positivos dado que mide la precisión en los eventos de la población.

Clasificación Binaria II

- De manera inversa, la especificidad es definida como el ratio de observaciones que no son los eventos que son predichos como no eventos (ratio de verdaderos negativos).

$$Especificidad(d) = \frac{TN}{FP + TN}$$

- La falsa alarma o ratio de falsos positivos es definido como uno menos la especificidad.

$$Falarm(d) = 1 - Especificidad(d) = \frac{FP}{FP + TN}$$

- La precisión es la comparación entre los verdaderos positivos con las instancias predichas como positivas:

$$Precision(d) = \frac{TP}{TP + FP}$$

Clasificación Binaria III

- Basado en la precisión y la sensibilidad, la medida-F (F1-score) puede ser usada, la cuál es la media armónica entre la precisión y la sensibilidad:

$$F_1(d) = 2\frac{precision * sensibilidad}{precision + sensibilidad}$$

Equilibrio entre sensibilidad y la especificidad I

- Asumiendo un nivel fijo de exactitud para el modelo, normalmente hay un equilibrio que debe hacerse entre la sensibilidad y la especificidad.
- Intuitivamente, incrementando la sensibilidad de un modelo es probable en incurrir en una pérdida de especificidad, debido a que más observaciones son predichas como eventos.
- Equilibrios potenciales entre la sensibilidad y especificidad pueden ser apropiados cuando hay diferentes penalidades asociadas con cada tipo de error. En el caso de filtro de spam, usualmente el foco es la especifidad, debido a que la mayoría de personas está dispuesta a tolerar algunos spams si es que los emails de familiares y colaboradores no van a ser borrados.

Equilibrio entre sensibilidad y la especificidad II

- Usualmente, es de interés in tener una medida única que refleja los ratios de falsos positivos y los de falsos negativos. El índice de Youden (Youden, 1950) definido como:

$$J = sensibilidad + Especificidad - 1$$

  mide las proporciones de observaciones correctamente predichas tanto para los eventos como para los no eventos. En algunos contextos esto puede ser un método apropiado para resumir la magnitud de ambos tipos de errores.
- La curva ROC (receiver operating characteristic) es una de las técnicas más comunes para evaluar la combinación de la sensibilidad y la especificidad dentro de un único valor.
- Un aspecto que se suele pasar por alto al evaluar la sensibilidad y la especificidad es que son medidas condicionales.

Equilibrio entre sensibilidad y la especificidad III

- La sensibilidad es el ratio de precisión solo para los eventos de la población (y la especificidad para los no eventos)
- Usando la sensibilidad y la especificidad una obstetriz puede hacer afirmaciones como "asumiendo que el feto no tenga síndrome de Down, la prueba tiene una precisión de 95 %".
- Sin embargo, estas afirmaciones pueden ser de poca ayuda para un paciente dado que, para una nueva observación, todo lo que se sabe es la predicción. La persona que usa un modelo predictivo está típicamente interesado en preguntas no condicionales tales como "¿Cuál es la posibilidad de que el feto tenga un desorden genético?".
- Esto depende de tres valores: la sensibilidad y la especifidad de la prueba diagnóstica y la prevalencia del evento en la población. De manera intuitiva, si el evento es raro, esto debería ser reflejado en la respuesta.

Equilibrio entre sensibilidad y la especificidad IV

- La prevalencia está definida como

$$Prevalencia = \frac{TP + FN}{TP + FN + FP + FN}$$

- Tomando la prevalencia en consideración, el análogo a la sensibilidad es el valor positivo pronosticado (PPV).

$$PPV = \frac{sensibilidad * Prevalencia}{(Sentitividad * Prevalencia) + ((1 - Especificidad) * (1 - Prevalencia))}$$

- El PPV responde a la pregunta "¿Cuál es la probabilidad de que esta observación sea un evento?"
- Del mismo modo, el análogo a la especificidad es el valor negativo pronosticado (NPV).

$$NPV = \frac{Especificidad * (1 - Prevalencia)}{(Prevalencia * (1 - sensibilidad) + (Especificidad * (1 - Prevalencia))}$$

Equilibrio entre sensibilidad y la especificidad V

- En relación a la estadística Bayesiana, la sensibilidad y la especificidad son las probabilidades condicionales, la prevalencia es la priori, y los valores pronosticados positivos/negativos las probabilidades a posteriori.

Selección de Medidas I

- No existe una respuesta sencilla a la pregunta relacionada a que medida de evaluación usar.
- En general, ningún clasificador es óptimo para todas las métricas de evaluación.
- Cuando se evalúa un problema de clasificación en general, la exactitud es más que suficiente, junto con el análisis del coeficiente Kappa de Cohen.
- Por supuesto, si existen problemas de categorías no balanceadas, uno debería tener en cuenta las medidas F para verificar si existe un buen balance entre la precisión y la sensibilidad.
- En problemas específicos, hay que tener mucho cuidado con la selección de la medida.

Selección de Medidas II

- Por ejemplo, cuando existe un alto costo relacionado a la clasificación de la clase negativa, una falsa alarma muy alta es problemática.
- En algunos casos es recomendable usar múltiples medidas, y buscar un balance entre ellas.