La evaluación es la clave para lograr un proceso real en la minería de datos. Hay diferentes métodos para inferir una estructura de datos, para determinar cuál es mejor usar, lo mejor es evaluar como funcionan y comparar unos con otros, sin embargo, no es tan sencillo, el rendimiento en el conjunto de entrenamiento no es un buen indicador de rendimiento en un conjunto de prueba independiente, necesitamos formas de predecir los limites de desempeño en la práctica. Algunas situaciones implican predecir la clase probabilidades en lugar de las clases mismas, y otras implican predecir valores numéricos en lugar de nominales. Se necesitan diferentes métodos en cada caso.

**Training and testing.**

Para problemas de clasificación, es natural medir el desempeño de un clasificador en términos de la tasa de error. El clasificador predice la clase de cada instancia: si es correcto, eso se cuenta como un éxito; si no, es un error. Pero no es un buen indicador en datos nuevos.

La tasa de error en los datos de entrenamiento se denomina error de resustitución porque es calculado resustituyendo las instancias de entrenamiento en un clasificador que fue construido a partir de ellos. Aunque no es un predictor confiable de la verdadera tasa de error sobre nuevos datos. Para predecir el desempeño de un clasificador en nuevos datos, necesitamos evaluar su tasa de error en un conjunto de datos que no jugó ningún papel en la formación del clasificador. Este conjunto de datos independiente se denomina conjunto de prueba. Es importante que los datos de prueba no se utilicen de ninguna manera para crear el clasificador.

Se suele hablar de tres conjuntos de datos: los datos de entrenamiento, los datos de validación y datos de prueba. Los datos de entrenamiento son utilizados por uno o más esquemas de aprendizaje para crear clasificadores. Los datos de validación se utilizan para optimizar los parámetros de esos clasificadores o para seleccionar uno en particular. Luego, los datos de prueba se utilizan para calcular la tasa de error del método final optimizado. Cada uno de los tres conjuntos debe ser elegido de forma independiente: el conjunto de validación debe ser diferente del conjunto de entrenamiento para obtener un buen desempeño en la etapa de optimización o selección, y el equipo de prueba debe ser diferente de ambos para obtener una estimación fiable de la tasa de error real.

El verdadero problema se produce cuando no se dispone de una gran cantidad de datos. En muchas situaciones, los datos de entrenamiento deben clasificarse manualmente, al igual que los datos de prueba, para obtener estimaciones de error. Esto limita la cantidad de datos que se pueden ser utilizados para el entrenamiento, validación y pruebas, y el problema se convierte en cómo hacer lo mejor con un conjunto de datos limitado.

**Predicting Performance.**

En estadística, una sucesión de eventos independientes que tienen éxito o fracasan se llama un proceso de Bernoulli. La media y la varianza de un único ensayo de Bernoulli con tasa de éxito p son p y p (1 - p), respectivamente. Si se toman N ensayos de un proceso de Bernoulli, la tasa de éxito esperada f = S / N es una variable aleatoria con la misma media p; la varianza es reducida por un factor de N a p (1 - p) / N. Para N grande, la distribución de esta variable aleatoria se acerca a la distribución normal. La probabilidad de que una variable aleatoria X, con media cero, se encuentre dentro de un cierto rango de confianza de ancho 2z es Pr [-z <= X <= z] = c

Para una distribución normal, los valores de c y los valores correspondientes de z se dan en tablas impresas en la parte posterior de la mayoría de los textos estadísticos. Sin embargo, las tabulaciones toman convencionalmente una forma ligeramente diferente: dan la confianza de que X estará fuera del rango, y lo dan solo para la parte superior del rango: Pr[X >= z]

Esto se llama probabilidad de una cola porque se refiere solo a la "cola" superior de la distribución. Las distribuciones normales son simétricas, por lo que las probabilidades de la cola inferior son iguales.

**Cross-Validation.**

El método de retención reserva una cierta cantidad para la prueba y usa el resto para entrenamiento es común mantener un tercio de los datos para probar y usar los otros dos tercios para el entrenamiento.

Se puede presentar un mal caso en el que la muestra utilizada para la capacitación (o las pruebas) no es representativa. No se puede saber si una muestra es representativa o no. Pero hay una simple comprobación que podría valer la pena hacer: cada clase del conjunto de datos completo debe estar representado aproximadamente en la proporción correcta en el entrenamiento y conjuntos de prueba. Este procedimiento se llama estratificación y podríamos hablar de resistencia estratificada.

Una forma más general de mitigar cualquier sesgo causado por la muestra particular elegida para el procedimiento es repetir todo el proceso, entrenamiento y prueba, varias veces con diferentes muestras aleatorias. Este es el método de retención repetido de la estimación de la tasa de error.

En la validación cruzada, decides un número fijo de pliegues, o particiones, de los datos. Supongamos que usamos tres. Luego, los datos se dividen en tres particiones aproximadamente iguales; cada una a su vez se utiliza para pruebas y el resto se utiliza para entrenamiento. Es decir, usas dos tercios de los datos para entrenamiento y un tercio para probar, y se repite el procedimiento tres veces hasta que al final, cada instancia se ha utilizado exactamente una vez para la prueba. El estándar es de usar diez particiones.

**Other Estimates.**

La validación cruzada de diez particiones es la forma estándar de medir la tasa de error de un esquema de aprendizaje en un conjunto de datos en particular. Pero en su lugar se utilizan muchos otros métodos. Dos que son particularmente frecuentes son Leave-one-out-cross-validation y el bootstrap.

**Leave-one-out-cross-validation.**

La validación cruzada de dejar uno fuera es simplemente una validación cruzada de n veces, donde n es el número de instancias en el conjunto de datos. Cada instancia, a su vez, se deja fuera, y el esquema de aprendizaje está entrenado en todas las instancias restantes. Se juzga por su exactitud en la instancia restante, uno o cero para el éxito o el fracaso, respectivamente. Los resultados de todos los n juicios, uno para cada miembro del conjunto de datos, se promedian, y el promedio representa la estimación del error final.

**The Bootstrap.**

Se basa en el procedimiento estadístico de muestreo con reposición. La idea del bootstrap es muestrear el conjunto de datos con reemplazo para formar un conjunto de entrenamiento. Las instancias se muestrean n veces, con reemplazo, para dar otro conjunto de datos de n instancias. Dado que algunos elementos de este segundo conjunto de datos se repetirán debe haber algunas instancias en el conjunto de datos original que no se han seleccionado; estas serán las instancias de prueba.

La cifra obtenida al entrenar un sistema de aprendizaje en el conjunto de entrenamiento y calcular su error sobre el conjunto de prueba será una estimación pesimista de la tasa de error real para compensar esto, combinamos la tasa de error del conjunto de prueba con el error de resustitución en las instancias del conjunto de entrenamiento. Luego, todo el procedimiento de arranque se repite varias veces, con diferentes muestras de reemplazo para el conjunto de entrenamiento, y los resultados se promedian.

**Comparing Data Mining Schemes.**

A menudo necesitamos comparar dos esquemas de aprendizaje diferentes sobre el mismo problema para vea cuál es el mejor para usar.

Parece simple: Estimar el error usando validación cruzada o cualquier otro procedimiento de estimación adecuado, quizás repitiendo varias veces, y elegir el esquema con la estimación más pequeña. Sin embargo, puede ser que la diferencia se deba simplemente a un error de estimación y, en algunos casos, circunstancias es importante determinar si un esquema es realmente mejor que otro sobre un problema particular.

**Predicting Probabilities.**

El resultado de cada instancia de prueba es correcto, si la predicción concuerda con el valor real para esa instancia, o incorrecta, si no lo hace. Si el esquema de aprendizaje, cuando se aplica realmente, da como resultado una predicción correcta o incorrecta, el éxito es la medida correcta para usar. Esto a veces se denomina función de pérdida 0 - 1: la "pérdida" es 0 si la predicción es correcta o 1 si no lo es.

La mayoría de los esquemas de aprendizaje pueden asociar una probabilidad con cada predicción (como lo hace el esquema Naïve Bayes). Podría ser más natural tener en cuenta esta probabilidad al juzgar la corrección.

**Quadratic Loss Function.**

Un criterio que se utiliza con frecuencia para evaluar la predicción probabilística es la función de pérdida cuadrática. la función de pérdida cuadrática obliga al predictor a ser honesto acerca de elegir su mejor estimación de las probabilidades o, más bien, da preferencia a predictores que pueden hacer la mejor suposición de las probabilidades verdaderas.

**Informational Loss Function.**

Un problema con la función de pérdida de información es que si asigna una probabilidad de 0 a un evento que realmente ocurre, el valor de la función es infinito. La función de pérdida de información no asigna probabilidad cero a ningún resultado. Esto hace conducir a un problema cuando no hay información disponible sobre ese resultado sobre el cual basar una predicción.

**Counting the cost.**

Las evaluaciones que se han discutido hasta ahora no tienen en cuenta el costo de tomar decisiones equivocadas o clasificaciones equivocadas. La precisión de la evaluación por clasificación asume tácitamente costos de error iguales. El costo de identificar erróneamente los problemas de una máquina que resulta estar libre de fallas es menor que el costo de pasar por alto los problemas de una que está a punto de fallar.

Los verdaderos positivos (TP) y los verdaderos negativos (TN) son clasificaciones correctas. Un falso positivo (FP) es cuando el resultado se predice incorrectamente como sí (o positivo) cuando en realidad es no (negativo). Un falso negativo (FN) es cuando el resultado se predice incorrectamente como negativo cuando en realidad es positivo. La tasa de verdaderos positivos es TP dividida por el número total de positivos, que es TP + FN; la tasa de falsos positivos es FP dividida por el número total de negativos, que es FP + TN.

La tasa de éxito general es el número de clasificaciones correctas dividido por el número total de clasificaciones. Finalmente, la tasa de error es 1 menos esto.

En la predicción multiclase, el resultado de un conjunto de prueba a menudo se muestra como una matriz de confusión bidimensional con una fila y una columna para cada clase. Los buenos resultados corresponden a grandes números por la diagonal principal y elementos pequeños, idealmente cero, fuera de la diagonal.

La estadística Kappa se utiliza para medir la concordancia entre lo previsto y las categorizaciones observadas de un conjunto de datos, mientras se corrige un acuerdo que ocurre por casualidad. Sin embargo, al igual que la tasa de éxito simple, no tiene en cuenta los costos.

**Cost-Sensitive Classification.**

Si se conocen los costos, se pueden incorporar a un análisis financiero del proceso de toma de decisiones. En el caso de dos clases, los costos se pueden resumir en forma de una Matriz de 2 × 2 en la que los elementos diagonales representan los dos tipos de clasificación correcta y los elementos fuera de la diagonal representan los dos tipos de error. En el caso multiclase esto se generaliza a una matriz cuadrada cuyo tamaño es el número de clases, y nuevamente los elementos diagonales representan el costo de la clasificación correcta.

Un análisis financiero completo del proceso de toma de decisiones también podría tener en cuenta el costo de usar la herramienta de aprendizaje automático, incluido el costo de recopilar los datos de entrenamiento, y el costo de usar el modelo, o la estructura de decisiones, incluido el costo de determinar los atributos para las instancias de prueba. Si se conocen todos los costos, y el número proyectado de los cuatro resultados diferentes en la matriz de costos se puede estimar, usando validación cruzada, es sencillo realizar este tipo de análisis financiero.

**Cost-Sensitive Learning.**

Los costos son ignorados en el momento del entrenamiento, pero utilizado en el momento de la predicción. Una alternativa es hacer solo el opuesto: Tenga en cuenta la matriz de costos durante el proceso de capacitación e ignore costos en el momento de la predicción. La idea es generar datos de entrenamiento con una proporción diferente de casos de sí y no.

El esquema que se esfuerza por minimizar el número de errores, elegirá una estructura que está sesgada hacia evitar errores en las instancias no porque tales errores se penalizan efectivamente diez veces. Si los datos con la proporción original sin instancias se utilizan para la prueba, se cometerán menos errores en estos que en con instancias, es decir, habrá menos falsos positivos que falsos negativos, porque los falsos positivos se han valorado diez veces más que los falsos negativos. Variar la proporción de instancias en el conjunto de entrenamiento es una técnica general para construir clasificadores sensibles a los costos.

**Lift Charts.**

En la práctica, los costos rara vez se conocen con algún grado de precisión, si conociera los costos, podría determinar la recompensa implícita en un factor de sustentación particular. Dado un esquema de aprendizaje que genera probabilidades para la clase predicha de cada miembro del conjunto de instancias de prueba (como lo hace Naïve Bayes), su trabajo es encontrar subconjuntos de casos de prueba que tienen una alta proporción de casos positivos, mayor que en el conjunto de prueba en su conjunto. Para hacer esto, las instancias deben ordenarse en orden descendente orden de probabilidad predicha de sí. Luego, para encontrar una muestra de un tamaño dado con la mayor proporción posible de casos positivos, solo lea el número requerido de instancias fuera de la lista, comenzando por la parte superior. Si se conoce la clase de cada instancia de prueba, puede calcular el factor de elevación simplemente contando el número de instancias positivas que incluye la muestra, dividiendo por el tamaño de la muestra para obtener una proporción de éxito, y dividiendo por la proporción de éxito para el conjunto de prueba completo para determinar el factor de elevación.

Si se conocieran los diferentes costos involucrados, podría calcularlos para cada muestra tamaño y elija el más rentable. Pero una descripción gráfica de las diversas posibilidades a menudo será mucho más reveladora que presentar una sola decisión óptima. Los valores de costo o beneficio asociados con clasificaciones incorrectas o correctas pueden ser ingresado en la matriz y afectar la forma de la curva en el gráfico.

**ROC Curves.**

Los gráficos de elevación son una herramienta valiosa, ampliamente utilizada en marketing están estrechamente relacionados a una técnica gráfica para evaluar esquemas de minería de datos conocida como curvas ROC. Las curvas ROC representan el desempeño de un clasificador sin tener en cuenta la distribución de clases o el costo de error. Trazan la tasa verdadera positiva en el eje vertical contra la verdadera tasa negativa tasa en el eje horizontal.

**Recall-Precision Curves.**

La gente ha lidiado con la compensación fundamental ilustrada por los gráficos de elevación y curvas ROC en una amplia variedad de dominios. La recuperación de información, por ejemplo, los expertos en recuperación de información utilizan curvas de precisión de recuperación que trazan uno contra el otro, para diferentes números de documentos recuperados, en solo de la misma manera que las curvas ROC y los gráficos de elevación, excepto que, debido a que los ejes son diferentes, las curvas son de forma hiperbólica y el punto de operación deseado está hacia la parte superior derecha.

**Cost Curves.**

Las curvas ROC y sus parientes son muy útiles para explorar las compensaciones entre diferentes clasificadores en una variedad de escenarios. Sin embargo, no son ideales para evaluar modelos de aprendizaje automático en situaciones con costos de error conocidos. Las curvas de costos son un tipo diferente de visualización en la que un solo clasificador corresponde a una línea recta que muestra cómo varía el rendimiento a medida que cambia la distribución de clases. Nuevamente, funcionan mejor en el caso de dos clases, aunque siempre puede convertir un problema multiclase en uno de dos clases seleccionando una clase y evaluándola con las restantes. Los buenos clasificadores tienen bajas tasas de error, por lo que desea estar lo más cerca posible de la parte inferior del diagrama.

Cada punto en un gráfico de elevación, curva ROC o curva de precisión de recuperación representa un clasificador, normalmente obtenido mediante el uso de diferentes valores de umbral para un método como Naïve Bayes. Las curvas de costos representan cada clasificador mediante una línea recta y una suite de clasificadores formara una curva cuyo límite inferior muestra qué tan bien ese tipo de clasificador puede funcionar si el parámetro está bien elegido.

**Evaluating numeric prediction.**

Los principios básicos: utilizar un conjunto de pruebas independientes en lugar del conjunto de entrenamiento para la evaluación del desempeño, el método, validación cruzada — se aplica igualmente bien a la predicción numérica.

Pero lo medida básica de calidad ofrecida por la tasa de error ya no es apropiada: los errores no son simplemente presente o ausente; vienen en diferentes tamaños.

El error cuadrático medio es la medida principal y más utilizada; a veces se toma la raíz cuadrada para darle las mismas dimensiones que el valor predicho sí mismo.

El error absoluto medio es una alternativa: simplemente promedie la magnitud de los errores individuales sin tener en cuenta su signo. El error cuadrático medio tiende a exagerar el efecto de los valores atípicos: instancias en las que el error de predicción es mayor que los demás, pero el error absoluto no tiene este efecto: se tratan todos los tamaños de error uniformemente según su magnitud.

El error relativo al cuadrado se refiere a algo bastante diferente. El error se hace en relación con lo que habría sido si se hubiera utilizado un predictor simple. El predictor simple en cuestión es solo el promedio de los valores reales del entrenamiento. Por lo tanto, el error cuadrático relativo toma el error cuadrático total y lo normaliza dividiendo por el error cuadrático total del predictor predeterminado.

La siguiente medida de error se conoce con el glorioso nombre de error absoluto relativo y es solo el error absoluto total, con el mismo tipo de normalización. En estas tres medidas de error relativo, los errores se normalizan por el error del predictor simple que predice valores medios.

La última medida es el coeficiente de correlación, que mide la correlación estadística entre las a y las p. Los rangos del coeficiente de correlación van desde 1 para resultados perfectamente correlacionados, pasando por 0 cuando no hay correlación, hasta –1 cuando los resultados están perfectamente correlacionados de forma negativa. Por supuesto, valores negativos no debería ocurrir para métodos de predicción razonables. La correlación es ligeramente diferente de las otras medidas porque es independiente de la escala en el sentido de que, si toma un conjunto particular de predicciones, el error no cambia si todas las predicciones se multiplican por un factor constante y los valores reales no se modifican.

**Minimum Description Length Principle.**

El principio de longitud mínima de descripción, o MDL, toma la postura de que la mejor teoría para un cuerpo de datos es aquella que minimiza el tamaño de la teoría más la cantidad de información necesaria para especificar las excepciones relativas a la teoría se aplica al aprendizaje automático de la siguiente manera: Dado un conjunto de casos, un esquema de aprendizaje infiere una teoría, por muy simple que sea la mejor teoría es la que minimiza el número de bits necesarios para comunicar la teoría, junto con la etiquetas de los ejemplos de los que se hizo.

**Applying The MDL Principle to Clustering.**

Una de las cosas buenas del principio de longitud mínima de descripción es que, a diferencia de otros criterios de evaluación, se puede aplicar en circunstancias muy diferentes. La agrupación parece intrínsecamente difícil de evaluar. Considerando que la clasificación o el aprendizaje de asociación tiene un criterio objetivo de éxito la agrupación se puede evaluar desde una perspectiva de longitud de descripción.

Suponga que una técnica de aprendizaje por grupos divide el conjunto de entrenamiento E en k grupos, si estos grupos son naturales, debería ser posible usarlos para codificar E más eficientemente, si los datos exhiben una agrupación extremadamente fuerte, esta técnica dará como resultado una menor longitud de descripción que simplemente transmitir los elementos de E Sin embargo, si el efecto de agrupamiento no es tan fuerte, es probable que aumente en lugar de reducir la longitud de la descripción.

La formulación MDL, aplicada correctamente, puede ser lo suficientemente flexible para soportar la evaluación de la agrupación.