DP2

Modeling

De el árbol de decisión se puede concluir cosas como que el tamaño de la familia es más importante que la clase de pasajero para las mujeres. En general es más difícil que un hombre sobreviva debido a la ley de hombres y mujeres primero. Usar decisión tree, naive bayes y rule induction permite visualizar como se diferencian los algoritmos en cuanto a detectar los patrones y generar una solución.

Scoring

El scoring se trata sobre generar predicciones en nuevos datos con un modelo generado anteriormente. En este caso se usó un modelo naive bayes entrenado con datos etiquetados y luego se probó su rendimiento con otros datos sin etiquetar. El resultado presenta un porcentaje de confianza para cada predicción tanto como para no o si.

Test Splits and Validation

Es importante medir bien el rendimiento de un modelo porque una mala medida puede ser fatal en un sistema. Obviamente si se testea el rendimiento de un modelo con los mismos datos con que fue entrenado este va a ser muy alto, pero falso, por ello existen formas de evitar esa mala práctica. En este tutorial lo que se hace es dividir el dataset en dos partes, una con el 70 y otra con el 30 por ciento de los datos. Esto presenta sus problemas, pero es mejor que testear con los mismos datos de entrenamiento. El resultado es una matriz que indica la precisión del modelo y en que predicciones falló.

Cross Validation

En este tutorial se usa una técnica de testeo de modelos que consta de dividir el dataset en n partes. Luego se selecciona una parte y se la usa para testear el modelo, dejando el entrenamiento para el resto del dataset. Luego se mide la performance de cada iteración y se promedia el error. La ventaja de esta forma de validación es que se usa todo el dataset para testing, usar solo una parte puede causar que no se mida bien el rendimiento ante la posibilidad de elegir datos sesgados con ciertas características. La desventaja de esta estrategia es que requiere más poder de cómputo.

Visual Model Comparision

La Receiver Operationg Characteristics muestra cuan bien funciona un modelo de machine learning binario. Se pueden crear múltiples curvas ROCs para comparar distintos modelos. La idea es mostrar la tasa de verdaderos positivos frente a la tasa de falsos positivos. Cuando el modelo es muy impreciso la curva se acerca a una diagonal desde el origen hacia el infinito positivo. En el tutorial se usan tres modelos distintos. Se puede ver en el resultado que la gráfica muestra al decisión tree más cercano a la esquina superior izquierda, luego lo sigue la rule induction y por último el modelo de naive bayes. En general todos se alejan de ser una línea diagonal.