



Aprendizaje de Máquina



Menú

- Funciones de pérdida
- Evaluación de modelos
 - Algunas medidas
 - Tipos de errores
 - Curva ROC
- Cómo escoger los parámetros de un modelo



Funciones de pérdida

- Estas funciones se utilizan como las medidas de error que guían el ajuste de un modelo
- Regresión
 - La suma de diferencias al cuadrado (norma L_2)
 - La suma de las diferencias absolutas (L_1)
- Clasificación
 - Cross-entropy $H(p, q) = - \sum_x p(x) \log q(x).$
 - Donde p y q son distribuciones de probabilidad, p es la real y q es la estimada
 - Norma L_2
- El objetivo de la función de pérdida es proveer a los modelos con la mayor información para que se ajusten de mejor manera



Funciones de pérdida

- Normalmente un modelo busca minimizar el promedio de estas medidas para los datos de entrenamiento (y de prueba)
- Las medidas de error que discutiremos a continuación tienen que ver con lo que se reporta acerca del desempeño final de un modelo

Evaluación de Modelos

- Regresión

- Por lo general la medida de error utilizada es el error cuadrático medio

$$Error_{Modelo}(Datos) = E((Modelo(X_i) - f(X_i))^2) = \frac{1}{N} \sum_{i=1}^N (Modelo(X_i) - f(X_i))^2$$

- Clasificación

- La medida de error esta dada por el número de predicciones incorrectas entre el número de predicciones totales

$$Error_{Modelo}(Datos) = \frac{1}{N} \sum_{i=1}^N I(Modelo(X_i) \neq f(X_i))$$

Donde N es el número de datos, $f(X_i)$ es el verdadero valor para el dato X_i y I es la función indicadora (vale 0 o 1)



Calidad de un Modelo

- Casi siempre necesitamos obtener una visión más fina del error para problemas de clasificación
 - Diferentes aplicaciones dan diferente importancia a en qué se equivoca el modelo
 - A cuántos clientes les doy mal servicio vs fraude que detecto
 - Cuántos créditos de alto monto apruebo vs cuantos declino
 - A cuánta gente le doy radiación innecesaria
 - A cuántos enfermos no les doy radiación

Error de Clasificación

Matriz de Confusión

- Muchas veces es útil dividir el desempeño del sistema con respecto a la clase o acción final en:
 - Verdaderos Positivos, Falsos Positivos, Verdaderos Negativos y Falsos Negativos
- Una manera común de visualizar esto es hacer una matriz en donde:
 - Cada renglón tiene el número de instancias de cada clase (según los ejemplos de entrenamiento, la clase real)
 - Cada columna tiene el número de instancias por clase según el clasificador (para cierto valor del umbral)



Matriz de Confusión

- Ejemplo de dos clases *si* y *no* :

Clasificamos como

		si no		}	Clasificación real	
}	si	no				
	3	1	si	}		
	2	6	no			

- En rojo están los errores
- Para k clases es una matriz de $k \times k$



Matriz de Confusión

- A partir de la matriz de confusión podemos derivar varias medidas de desempeño. Una medida común es para cada una de las i clases o categorías calcular:
 - Verdaderos positivos (Tp)
 - El número de instancias que clasificamos como de la categoría i que verdaderamente pertenecen a i
 - Falsos positivos (Fp)
 - El número de instancias que clasificamos como de la categoría i que verdaderamente pertenecen a otra categoría distinta de i
 - ¿Cuál es la clase positiva y cuál la negativa? Por lo general se etiqueta como positiva la que demanda una acción (dar radiación, declinar transacción,...) y/o la clase que tiene menos instancias



Matriz de Confusión

- Del ejemplo anterior, del los 12 ejemplos
 - Tp
 - De los 4 ejemplos de la categoría *sí*, el modelo identifica 3. La proporción es:
 - $Tp=3/4=0.75$
 - Fp
 - De los 8 ejemplos de la categoría *no* , clasificamos 2 como *sí*. La proporción es:
 - $Fp=2/8=0.25$



Matriz de Confusión

- TN
 - De los 8 ejemplos de la categoría *no* el modelo identifica 6. La proporción es:
 - $TN = 6/8 = 0.75$
- FN
 - De los 4 ejemplos de la categoría *sí*, el modelo falla en 1. La proporción es:
 - $FN = 1/4 = 0.25$



Matriz de Confusión

- Para más de dos categorías tenemos una matriz de $k \times k$ (k el número de categorías)
 - La entrada i, j contiene el número de instancias pertenecientes a la categoría i pero que fueron clasificadas como pertenecientes a j
 - En este caso los falsos positivos son la suma de todos los elementos clasificados como i que pertenecen a una categoría distinta
 - Los falsos negativos son todos los elementos de la clase i que con clasificados como de otra clase



Otras Medidas de Bondad

- Accuracy
 - $(TP+TN)/(TP+TN+FP+FN)$
- Precision
 - $TP/(TP+FP)$
- Recall
 - $TP/(TP+FN)$
- Muchas veces es importante contar con un solo número para poder optimizar el modelo
 - F-measure
 - $2*(Precision*Recall)/(Precision+Recall)$
 - Área bajo la curva ROC
- Entre otras.....
- Muchas veces es necesario crear medidas relevantes para el problema (e.g. dinero ahorrado,...)



Sensibilidad del Modelo

Umbralización

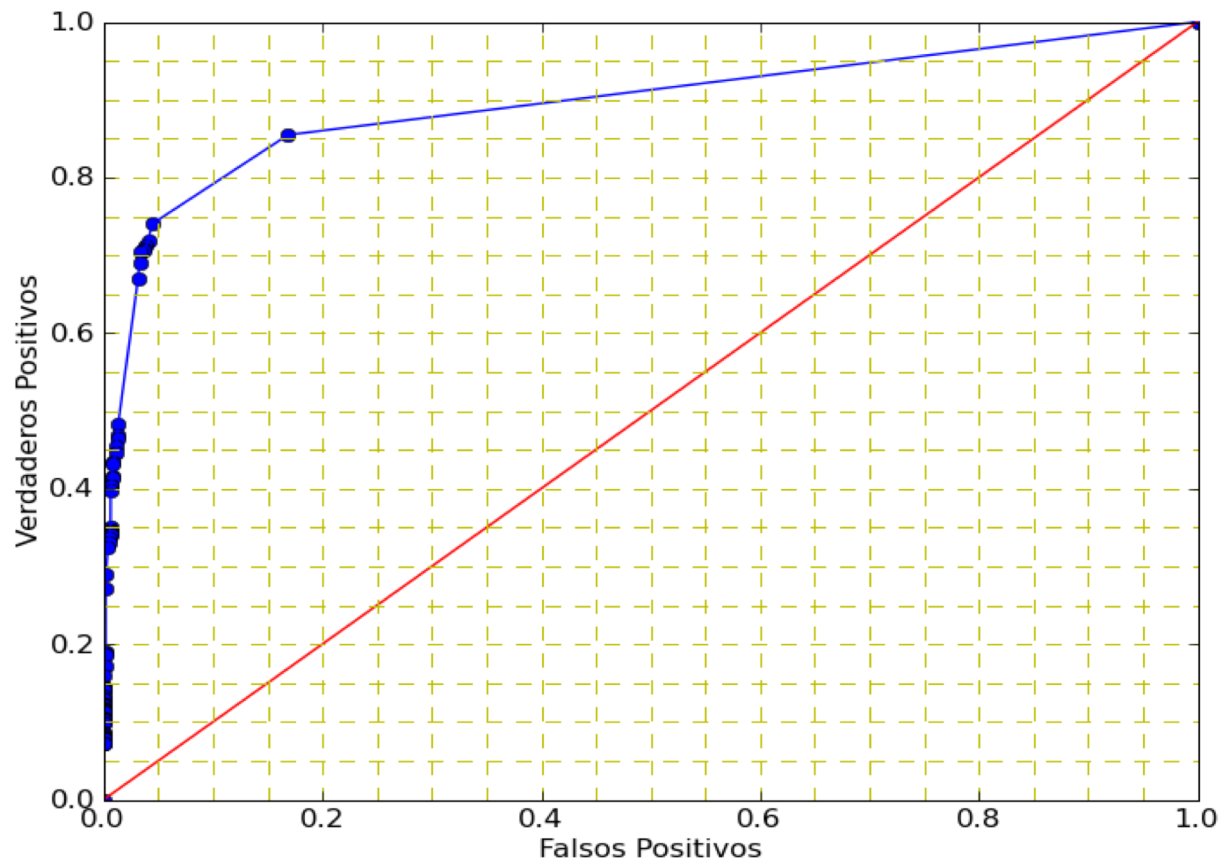
- En ocasiones los modelos de clasificación dan una calificación (o probabilidad) de pertenencia a una clase y por tanto la pertenencia de clase depende de un punto de corte (de la umbralización)
- Por ejemplo
 - En la detección de fraudes por lo general se asigna una calificación entre cero y uno a cada transacción. El operador del sistema debe de decidir a partir de que valor se considera algo como fraude
 - En el caso de detección de fraude se debe definir a partir de que “probabilidad” se recomienda tratamiento
- Para cada umbral, entonces, se calcula la bondad del modelo



Sensibilidad del Modelo

- Para examinar el desempeño del modelo en cuanto a su sensibilidad se utiliza una curva ROC (Receiver Operating Characteristic)
 - El eje de las x representa el porcentaje (o proporción) de FPs y el eje de las y el porcentaje de TPs
 - Cada punto en el gráfico representa la proporción FPs y TPs para una calificación dada. Notese que es acumulativo.
- En base a esto podemos escoger el umbral

Curva ROC





Sensibilidad del Modelo

- Los paquetes que reportan una matriz de confusión reportan el desempeño en el punto óptimo del ROC
 - Óptimo desde el punto de vista de alguna medida de error no necesariamente de lo que importa al negocio
- Es importante enfatizar que la importancia del tipo de error (FP o FN) depende de la aplicación y esto debe incluirse en la evaluación del método
 - Detección de spam
 - Detección de desperfectos en maquinaria



Ejercicio Opcional

- Para los datos EjercicioROC.csv
- Genere una curva de ROC en Excel
- Calcule el punto de corte óptimo en cuanto a asertividad (accuracy) y en cuanto a precisión
- Repita el ejercicio pero usando sklearn de python con los paquetes
 - roc_curve
 - Calcule el área bajo la curva