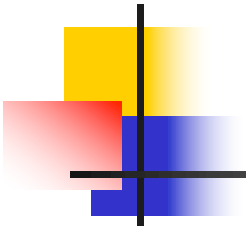
Aprendizaje de Máquina

ITAM



Menú

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

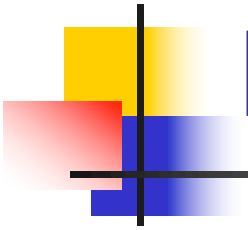
Funciones de pérdida

- Estas funciones se utilizan como las medidas de error que guían el ajuste de un modelo
- Regresión
  - Las 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 el ls 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
  - Algunas se usan también como funciones de pérdida para

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 un 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
- Un 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
si	3	1
no	2	6

Clasificación real

- En rojo están los errores
- Para k clases es una matriz de k X k

Matriz de Confusión

- A partir de la matriz de confusión podemos derivar varias medidas de desempeño. Un 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 *si*, el modelo identifica 3. La proporción es:
    - Tp=3/4=0.75
  - Fp
    - De los 8 ejemplos de la categoría *no ,* clasificamos 2 como *si.* 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 *si,* 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 X 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 elementes 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,…)

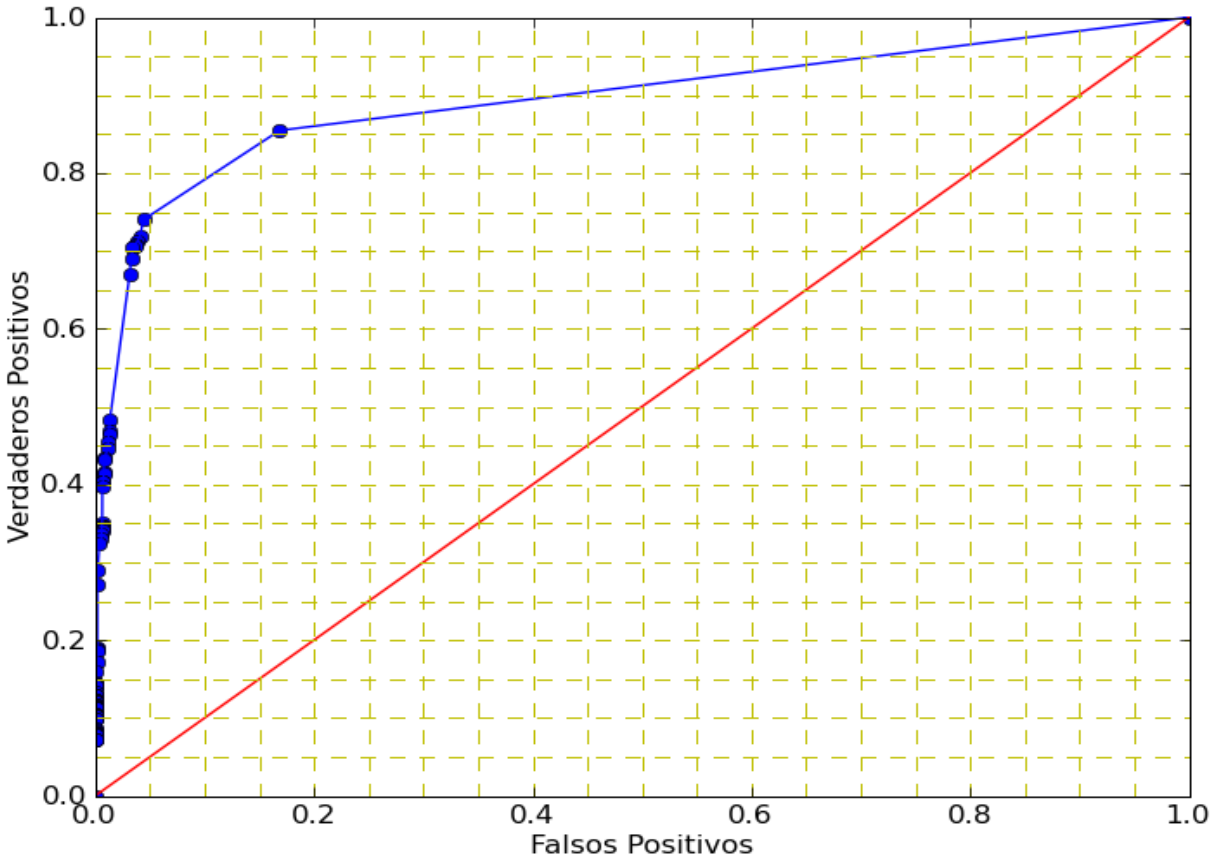
Sensibilidad del Modelo Umbralización

- En ocasiones los modelos de clasifiación 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

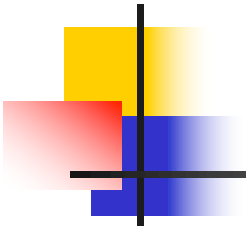
- Para examinar el desempeño del modelo en cuanto a su sensibilidad se utiliza una curva ROC (Receiver Operating Characteristic)
  - El eje de las x representa el porcentaje (o proporción) de FPs y el eje de las y el porcentaje de TPs
  - Cada punto en el gráfico representa la proporción FPs y TPs para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

Curva ROC


1.0
0.8
0.6
0.4
0.2
0.0
Verdaderos Positivos
0.0
0.2
0.4
0.6
0.8
1.0
Falsos Positivos

Sensibilidad del Modelo

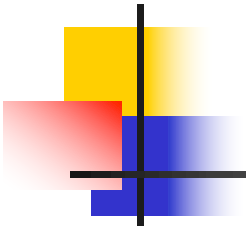
- Los paquetes que reportan una matriz de confusión reportan el desempeño en el punto óptimo del ROC
  - Óptimo desde el punto de vista de alguna medida de error no necesariamente de lo que importa al negocio
- Es importante enfatizar que la importancia del tipo de error (FP o FN) depende de la aplicación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



Ejercicio

- 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

Descomposición del Error



El Error

- El error de aprendizaje puede dividirse en tres componentes
  - El error irreductible dado a ruido
  - El error debido al sesgo del modelo. Lo que el modelo no puede capturar de la realidad
  - El error dado a alta varianza. Lo que el modelo captura pero no es real, es solo accidental en los datos de entrenamiento
- El alto sesgo se manifiesta como bajo-ajuste y la alta varianza como sobre-ajuste

Descomposición del Error en Sesgo y Varianza

- Supongamos que la función real es de la forma:
  - $y=f(x) + \varepsilon$, de ε es el ruido que se distribuye normalmente con media cero y varianza σ^2
- Nuestro modelo produce una predicción
  - $\hat{V}(x)$, para toda x
- Medimos el error como
  - $\Sigma(y-\hat{V}(x))^2$, en el caso de regresión (o de clasificación probabilística)

Descomposición del Error en Sesgo y Varianza

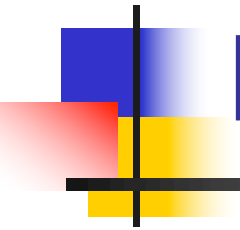
- Queremos estimar el error esperado para un nuevo punto x*

$$Err(x^*)=E[(y-\hat{V}(x^*))^2]$$
$$=E[(f(x^*)+\varepsilon-\hat{V}(x^*))^2]$$
$$=\sigma^2+[E(\hat{V}(x^*))-f(x^*)]^2+E[\hat{V}(x^*)-E(\hat{V}(x^*))]^2$$
$$=\sigma^2+Bias^2(\hat{V}(x^*))+Var(\hat{V}(x^*))$$
$$=ErrorIrreductible+Sesgo^2+Varianza$$

- Normalmente hay un compromiso entre sesgo y varianza

Descomposición del Error en Sesgo y Varianza

- Nótese que estas esperanzas son sobre todo lo que es aleatorio
  - Pesos iniciales (w's iniciales)
  - El conjunto de datos de entrenamiento (es la esperanza estimada sobre todos los posibles conjuntos de entrenamiento)
    - Por ejemplo para una regresion lineal: E(V^(x*)) es la salida esperada del modelo sobre todos los posibles conjuntos de entrenamiento con todas las posibles w's iniciales



Derivación
Descomposición de Error

Derivación Versión 1

- Una propiedad importante (truco para derivar)
  - $Var(X)=E(X^2)-[E(X)]^2$
- Sustituimos la variable aleatoria X por la discrepancia de nuestro modelo
  - $Var(\hat{V}(x)-f(x)-\varepsilon)=Var(\hat{V}(x)) + \sigma^2$
    - Porque la varianza de f(x) es cero pues no es una variable aleatoria y la covarianza entre el ruido y V^(x) es cero

Derivación Versión 1

- De la fórmula de la varianza sustituyendo:
- $Var(\hat{V}(x)) + \sigma^2 = E[(\hat{V}(x)-f(x)-\varepsilon)^2]-(E[\hat{V}(x)-f(x)-\varepsilon])^2$
  - $E[(\hat{V}(x)-f(x)-\varepsilon)^2]=MSE$ (error cuadrático medio)
  - $(E[\hat{V}(x)-f(x)-\varepsilon])^2=(E(\hat{V}(x))-E(f(x))-E(\varepsilon))^2$

  $=[E(\hat{V}(x))-f(x)]^2 =Bias^2$
    - Porque $E(f(x))= f(x)$ y $E(\varepsilon) =0$
- Sustituyendo en la primer formula
- $Var(\hat{V}(x)) + \sigma^2=MSE -Bias^2$
- $MSE=Var(\hat{V}(x)) + \sigma^2 + Bias^2$

Version 2
Derivación

- Algunas propiedades importantes:
  1. $E(E(x))=E(x)$
  2. $E((x-E(x))^2)=E[x^2-2xE(x)-E(x)^2]$

     $=E(x^2)-2E(xE(x))+E[E(x)^2]$

     $=E(x^2)-2E(x)E(x)+E(x)^2$

     $=E(x^2)-E(x)^2$
  3. $E(x^2)=E((x-E(x))^2)+E(x)^2$ (fórmula varianza)
  4. $E((c+N(0,\sigma))x)$

     $=E(cx+xN(0,\sigma))=cE(x)$ (la covarianza es cero)

Derivación

- Regresando al error esperado:

$E[(y-\hat{V}(x^*))^2] = E[y^2-2y\hat{V}(x^*)+\hat{V}(x^*)^2]$

$=E[y^2] -2E(y\hat{V}(x^*)) +E[\hat{V}(x^*)^2]$

$=E((y-E(y))^2) +E(y)^2$ (propiedad 3)

$-2E(\hat{V}(x^*))f(x^*)$ (propiedad 4)

$+ E[(\hat{V}(x^*)-E(\hat{V}(x^*)))^2] +E(\hat{V}(x^*))^2$ (propiedad 3)

$=E((y-E(y))^2)$ (ruido. El desarrollo da σ^2 usando prop.4 y 1)

$+E(y)^2 - 2E(\hat{V}(x^*))f(x^*) +E(\hat{V}(x^*))^2$ ($sesgo^2$ esto se reduce a $(y-E(\hat{V}(x^*)))^2$ note que $E(y)=y$)

$+E[(\hat{V}(x^*)-E(\hat{V}(x^*)))^2]$ (varianza)