Aprendizaje de Máquina

ITAM

Outline

- Loss functions
- Model evaluation
  - Some metrics
  - Types of errors
  - ROC curve

Loss Functions

- These are the functions used by the algorithms to evaluate the models' error and that guide the optimization

- Regression
  - Sum of the squared differences (L_2 Norm)

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

  - Sum of the absolute differences (L_1 Norm)

Loss Functions

- Classification
  - Total error: The proportion of misclassified examples

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

Where N is the number of examples, $f(X_i)$ is the true value for X_i and I is an indicator function (takes values of 0 or 1)

  - Cross-entropy $H(p,q) = -\sum_x p(x)\,\log q(x).$
    - Where p and q are probability distributions. p is the real one and q the estimate
  - L_2 norm whern the output is treated as a probability

Model quality

- Almost always we need to have a finer vision of the error for classification problems
  - Different applications give different importance to what the model is misclassifying
    - How many clients am I giving poor service vs how much fraud Im detecting
    - How many credits of high value I approve vs how many I decline
    - To how many patients am I administering radiation unnecessarily
    - To how many patients am I not giving radiation that need it

Classification Error
Confusion Matrix

- It is often useful to analize the model's performance in terms of:
  - True Positives, False Positives, True Negatives and False Negatives
- A common way to represent this is with a matrix where:
  - Each row has the number of instances of each class according to the their **true** label
  - Each column has the number of instances of each class according to the their **predicted** label

Confusion Matrix

- Two class example *si* y *no* :

Model classifies as:

si	no	
3	1	si
2	6	no

True class

- Mistakes in red
- For k clases we'll have a k X k matrix

Confusion Matrix

- Two class example *si* y *no* :

Model classifies as:

si	no	
TP	FN	si
FP	TN	no

True class

- Which class is positive and which negative? In general we say the postive class is the one that requires an action or that which has the least amount of instances (fraud in the case of fraud detection, cancer in the case of medical diagnosis)

Some Performace Measures

- Accuracy
  - (TP+TN)/(TP+TN+FP+FN)
- Precision
  - TP/(TP+FP)
- Recall (Tpr)
  - TP/(TP+FN)
- Fpr (False positive rate)
  - FP/(FP+TN)
- Often we need one value to compare models
  - F-measure
    - 2*(Precision*Recall)/(Precision+Recall)
  - Area under the ROC
- Among other measuers…..
- Almost always we need to create problem specific measures for instance “money saved”

Some examples

- Form the previous confusion matrix, del los 12 ejemplos
  - Recall (Tpr)
    - From the 4 *si* instances the model identifies 3
    - Tp=3/4=0.75
  - Fpr
    - From the 8 *no* examples the model misclassifies 2
    - Fp=2/8=0.25
  - Precision
    - From the 5 instances classified as *si*, 3 are correct
    - Precision=3/5
  - Accuracy
    - Proportion of correct classifications
    - Acccuracy=9/12

Confusion Matrix

- For k categories we'll have a k X k matrix
  - Entry *i,j* has the number of instances that belong to category *i* that are classified as category *j*

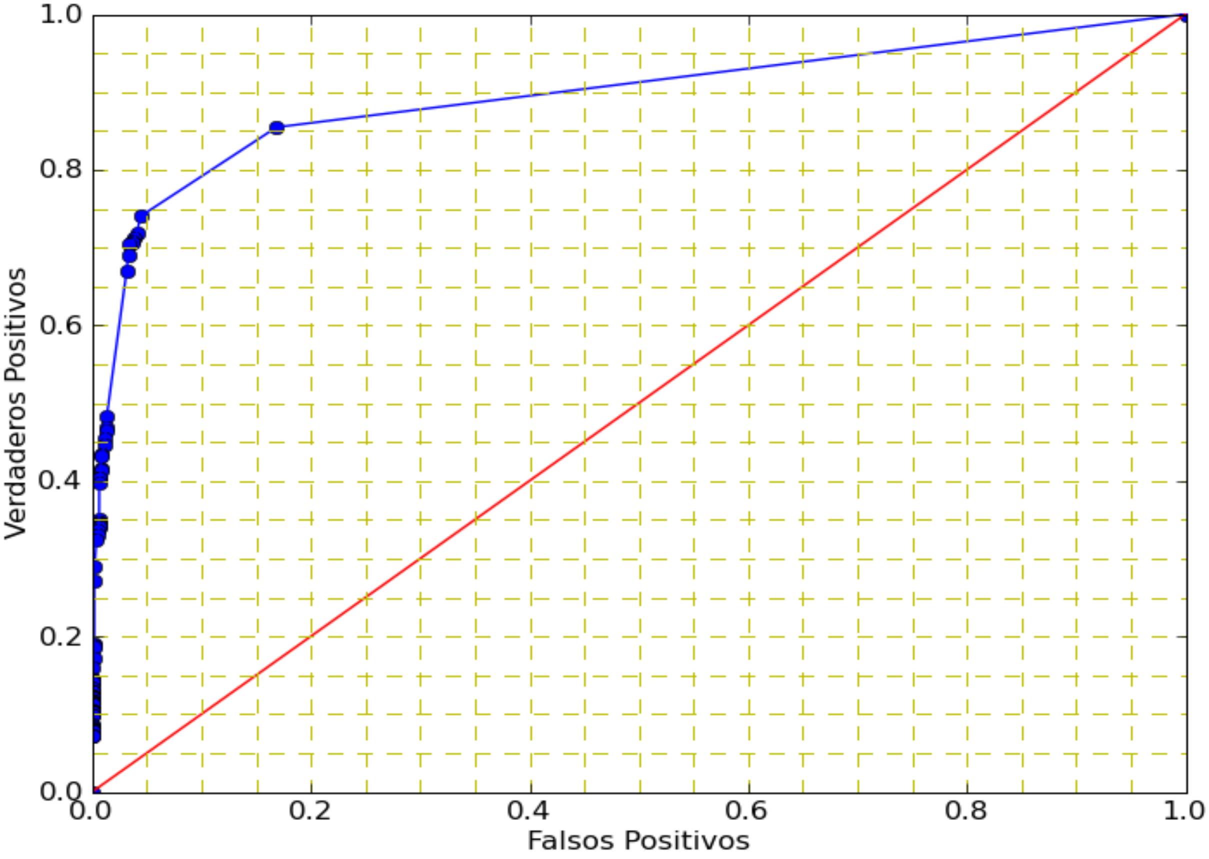
Model Sensitivity Thresholding

- On ocasión classification models ouput the probability of an instance belonging to a category or class rather that the category itself. Therefore we need to define a threshold value for assigning classes (with the spam filter example we set it at 0.5)
- For example
  - False postives are expensive in spam filters since too many false alerts cause people to stop using it. Perhaps in this case a higher threshold is warranted
- We calculate the performace metrics for several values of the threshold

Model Sensitivity

- In order to examine the sensitivity of the model and as a method for choosing a threshold we use the ROC (Receiver Operating Characteristic)
  - The x axis represents the false postive rate (Fpr). The y axis the Tpr
  - Every point on the plot represent the proportion of FPs and TPs using a fixed score as threshold This representation allows us to visualize the tradoffs of our model and choose a threshold

ROC



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

Model Sensitivity

- Many pakages automatically choose a threshold based on the best Tpr and Fpr tradeoff
  - This might not be the best choice for some applications
- Its important to emphasize that not all aplications are afected equally by FPs and FNs
  - Spam detection
  - Tumor detection

Exercise

- Data: EjercicioROC.csv
- Use a spreadsheet to plot a ROC
- What is the optimal threshold for accuracy? What is it for precision?
- Repeat using the packages (roc_curve) provided by sklearn
  - Compute the area under the curve

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
  - $V\hat{}(x)$, para toda x
- Medimos el error como
  - $\Sigma(y-V\hat{}\,(x))^{2},$ 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
  - $Var(V\hat{}(x)-f(x)-\varepsilon)=Var(V\hat{}(x)) + \sigma^2$
    - 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:
- $Var(V\hat{}(x)) + \sigma^2 = E[(V\hat{}(x)-f(x)-\varepsilon)^2]-(E[V\hat{}(x)-f(x)-\varepsilon])^2$
  - $E[(V\hat{}(x)-f(x)-\varepsilon)^2]=MSE$ (error cuadrático medio)
  - $(E[V\hat{}(x)-f(x)-\varepsilon])^2=(E(V\hat{}(x))-E(f(x))-E(\varepsilon))^2$

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

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:

$E[(y-V^\wedge(x^*))^2] = E[y^2 - 2yV^\wedge(x^*) + V^\wedge(x^*)^2]$

$= E[y^2] - 2E(yV^\wedge(x^*)) + E[V^\wedge(x^*)^2]$

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

$-2E(V^\wedge(x^*))f(x^*)$ (propiedad 4)

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

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

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

$+E[(V^\wedge(x^*)-E(V^\wedge(x^*)))^2]$ (varianza)