Machine Learning para el análisis de texto

Daniel Jiménez M.

Universidad Nacional de Colombia

16 -11 -2020

Bibliografía

- Brett Lantz, Machine Learning with R, 2016, PACKT
- Bradley B & Brandon G, Hands-On Machine Learning with R,2020, The R Series.
- Silge G, Kuhn M,Tidy Modeling with R,2020,The R Series.
- Silge G,Hvitfeldt E, Supervised Machine Learning for Text Analysis in R, 2020, The R Series.

Librerías

Las librerías para esta clase son :

```
library(tidyverse)
library(tidymodels)
library(skimr)
library(themis)
```

Algunas definiciones importantes

La ciencia de datos, el Machine Learning y la AI son areas diferentes de estudio y trabajo, a pesar que se ha popularizado el termino **Data Scientist** como una sumatoria de los anteriores y por fines técnicos esto se mantendrá así. Más a continuación se presentarán algunas definiciones importantes y el contexto en el cual se desenvuelven.

Algunas definiciones importantes

La siguiente imagen es una tesis de Robinson (2017) sobre las diferencias en acciones de los data scientist:

- **Data science** produces **insights**
- **Machine learning** produces **predictions**
- **Artificial intelligence** produces **actions**

Data Science

El foco principal de la ciencia de datos es la producción de insigths, esto se logra de la siguiente manera :

- Los insigths son entendibles ;
- Se basan en el entendimiento descriptivo y prescriptivo de un problema;
- Se intenta entender las causalidades entre las variables exploratorias;

Una definición clásica dice que el data science combina : Estadística, ingeniería de sistemas y conocimiento de negocios.

Data Science

Pero a diferencia de la definición clásica que es la que proveé al nombre de Data Scientist a la persona que hace un poco de todo, es la siguiente:

Data Science

El data science es la disciplina que envuelve las siguientes corrientes:

- Inferencia estadística;
- Visualización de datos;
- Diseño de experimentos;
- Dominio del negocio y lo más importante;
- Comunicación.

Machine Learning

El machine learning produce predicciones (llamese forecast o clasificación, detección de anomalías entre otros).

Se divide en :

- Análisis supervisado: los datos tienen la forma de $\{(x_1, y_1), ..., (x_n, y_n)\}$
  - El objetivo es estudiar el comportamiento de la variable **y**, condicional a las variables respuestas.
  - Puntualmente se debe entender el comportamiento (distribución) de la variable **y** dada las explicativas.

Machine Learning

- Análisis no supervisado: Los datos tiene la siguiente forma $\{x_1, ...x_n\}$
  - El objetivo es estudiar la variable **x** y los posibles conglomerados que se encuentren en ello.
  - Matemáticamente hay que estudiar la variable **x**

Inteligencia Artificial

Son acciones, desarrolladas por algoritmos que entienden la lógica del funcionamiento de un entorno.

Una inteligencia artificial que todos usamos (o eso espero) es google maps, cuando selecciona la mejor ruta!

Machine Learning para la clasificación de textos

En esta sección veremos dos ejemplos, uno será a nivel de clasificación, donde hablaremos de las principales métricas para evaluar estos modelos.

Algunos de los algoritmos de clasificación son:

- Decision tree;
- Naive Bayes;
- Knn;
- Logistic Regression;
- SVM;
- Linear Discrimination Analysis;
- Random Forest

En esta oportunidad nos concentraremos en la regresión logística y el random forest

Regresión Logística

- Es un modelo basado en variables categóricas;
- Sus variables explicativas se dan en valores continuos
- No solo clasifica sino que genera probabilidad sobre los eventos

Regresión Logística

Condiciones para un problema logístico

- Un problema binario;
- Cuando se necesite la probabilidad de una predicción;
- Cuando éxita un perímeto de decisión;
- Cuando se necesite saber el impacto de las features.

Regresión Logística

La función logística funciona a través de la sigmoide

Esta es la sigmoide

$$\sigma(\theta^t X) = \frac{1}{1 + e^{-\theta^t X}}$$

Notese que si $\theta^t X$ es muy grande entonces la función tiende a uno, en caso contrario a cero

Regresión Logística

###Proceso de la logística

- Se inicia en σ
- Se calcula $\hat{y} = \sigma(\theta^t X)$
- Se calculan los errores $y - \hat{y}$
- Se minimiza la función de costo

Regresión Logística

Función de costo : Mejora los parámetros de la regresión a través de interacciones del siguiente tipo

$$\sigma(w^T x + b) = \frac{1}{1 + e^{-(w^T x + b)}}$$

Donde

$$J = -\frac{1}{m} \sum_{i=1}^{m} [y^{(i)} \log(a^{(i)}) + (1 - y^{(i)}) \log(1 - a^{(i)})]$$

cambiando así los pesos de los parámetros

Regresión Logística

Para mejorar la función de costo es necesario el gradiente descendiente

Gradiente descendiente

Función que interactivamente busca encontrar la aproxiumación de σ que minimice la función de costo.

Métricas de clasificación

- Precisión : Calidad del modelo

$$\frac{TP}{TP + FP}$$

Esto se puede leer como la precisión del modelo

- Recall: Porcentaje de positivos que tiene el modelo

$$\frac{TP}{TP + FN}$$

Métricas de clasificación

- F1-Score: Mide el rendimiento del modelo, basado en la siguiente relación

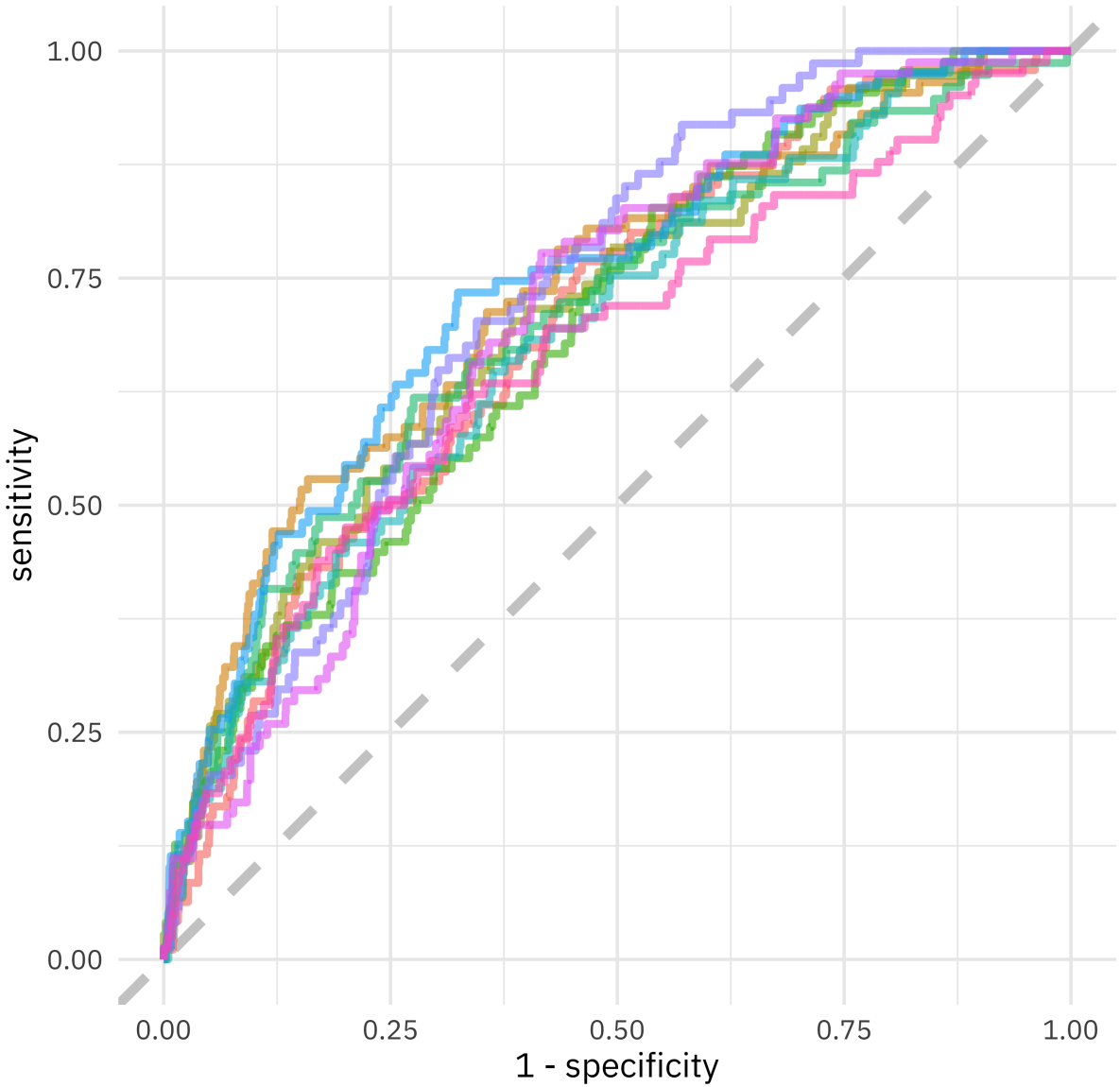
$$\frac{2 * (precision * recall)}{precision + recall}$$

Entre esta medida este más cercana a uno es mejor

- Accuracy : Mide el porcentaje de aciertos que tiene el modelo

$$\frac{(Tp + Tn)}{(Tp + Tn) + (Fp + Fn)}$$

Ejemplo


1.00
0.75
0.50
0.25
0.00
sensitivity
0.00
0.25
0.50
0.75
1.00
1 - specificity

Random Forest

Es una adaptación de los árboles de decisión, en donde lo valores de cada uno de sus árboles es independiente a nivel de interacción y dependiente de output del anterior.

Random Forest

- Es difícil de interpretar a nivel de sálidas
- Es útil cuando los modelos primarios no tienen suficiente accuracy
- Su ventaja se basa en la aleatoriedad.

Random Forest

- Su ventaja primordial es que si se tiene un training set grande, el valor de sus prediciones será buenismas
- Computacionalmente no es pesado
- Es super útil si se trabaja para estimar datos perdidos
- Posee buenos métodos experimentales

Observación : Tiende a sobre-ajustarse en tareas repetitivas de clasificación!!!!.