INFOX
educación sin fronteras

Clase 5: Modelos de clasificación

Mg. Gloria Rivas

Agenda

1. **Logistic Regression**
2. **K nearest neighbors**
3. **Naive Bayes**

Modelos de clasificación

- En el clásico modelo de regresión lineal se asume que la respuesta de la variable Y es cuantitativa, pero mucha veces la respuesta es cualitativa. Por ejemplo, el color de ojos si tomamos como colores azul, verde o negro.

- Estas variables cualitativas muchas veces son referidas como categóricas.

- En esta clase, vamos a aprender a predecir variables cualitativas que llamamos modelos de clasificación.

- Predecir una respuesta cualitativa para una observación puede ser referida como clasificación, dado que a la observación se le asigna a una categoría o clase.

- Por otro lado, muchas veces estos métodos de clasificación primero predicen la probabilidad de cada categoría como variable cualitativa.

- En esta clase veremos las más conocidas técnicas como Logistic reression, KNN y decision trees.

Modelos de clasificación

- Problemas de clasificación ocurren muy a menudo, incluso más que los problemas de regresión. Por ejemplo:

1. A person arrives at the emergency room with a set of symptoms that could possibly be attributed to one of three medical conditions. Which of the three conditions does the individual have?

2. An online banking service must be able to determine whether or not a transaction being performed on the site is fraudulent, on the basis of the user's IP address, past transaction history, and so forth.

3. On the basis of DNA sequence data for a number of patients with and without a given disease, a biologist would like to figure out which DNA mutations are deleterious (disease-causing) and which are not.

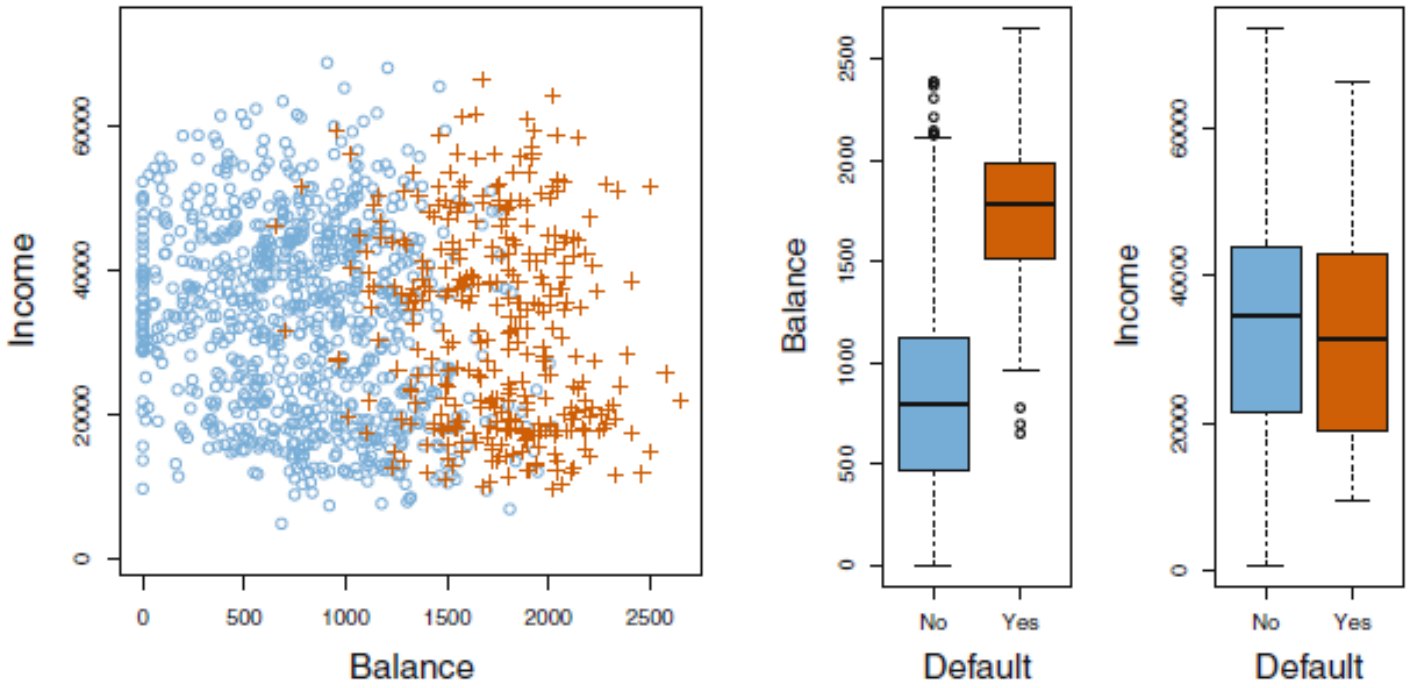
Modelos de clasificación

- Así como en problemas de regresión, en los modelos de clasificación tenemos un training set que podemos usar para construir a nuestro clasificador. Tenemos que tener en cuenta que este clasificador tiene que tener un buen performance no solo en el training set sino también en el test set.

- Vamos a ilustrar el ejemplo de clasificación usando un Dataset de un banco. Estamos interesados en predecir si un individuo va a ser default en su tarjeta de crédito tomando en cuenta su ingreso y su consumo mensual.

Modelos de clasificación

- En esta figura observamos el ingreso anual y el balance de tarjeta de crédito para una muestra de 10 000 clientes



- Parece que las personas que tienden a hacer default en su tarjeta son las que tienden a tener un mayor consumo.

- En este capítulo, vamos a aprender a hacer modelos para predecir el default (Y) tomando en cuenta el consumo (X1) y el ingreso (X2).

Modelos de clasificación

- Dado que Y no es una variable cuantitativa, una regresión lineal no es apropiada.

Por qué no una regresión lineal?

- Supongamos que estamos tratando de predecir la condición médica de un paciente en una sala de emergencia basado en sus síntomas. En un ejemplo simple, hay 3 posibles diagnósticos: stroke, drug, overdose y epileptic seizure.

- Consideremos estos valores como respuestas cuantitativas de la variable Y:

$$Y = \begin{cases} 1 & \text{if stroke;} \\ 2 & \text{if drug overdose;} \\ 3 & \text{if epileptic seizure.} \end{cases}$$

- Usando estos códigos, MCO puede ser usado para realizar un modelo de regresión lineal. Sin embargo, esto implica ordenar los códigos, poniendo drug overdose entre stroke y epileptic seizure, asumiendo que las diferencias entre stroke y drug overdose es la misma que drug overdose y epileptic seizure.

Modelos de clasificación

- En práctica no debemos asumir eso, por ejemplo podríamos elegir este código.

$$Y = \begin{cases} 1 & \text{if } \texttt{epileptic seizure}; \\ 2 & \text{if } \texttt{stroke}; \\ 3 & \text{if } \texttt{drug overdose}. \end{cases}$$

- Lo cual implicaría una relación completamente diferente a la anterior. Dependiendo de la codificación de los números llevaría a diferentes modelos lineales y por ende a diferentes sets de predicciones.

- Si la respuesta tuviera un orden natural como leve, moderado y grave entonces 1,2 y 3 como código sería razonable. Desafortunadamente, en general no hay una forma natural de convertir una respuesta cualitativa de más de dos niveles en una respuesta cuantitativa.

Modelos de clasificación

- Para una respuesta binaria, la situación es mucho mejor. Por ejemplo, solo habrían dos respuestas para un paciente stroke y drug overdose, entonces podríamos usar una dummy para codificar la respuesta.

$$Y = \begin{cases} 0 & \text{if stroke;} \\ 1 & \text{if drug overdose.} \end{cases}$$

- De esa manera, podríamos usar una regresión lineal y predecir drug overdose si Y > 0.5 y stroke de otra manera.

- Para respuestas binarias con código 0 y 1, regresiones por MCO hacen sentido. Sin embargo, nuestras estimaciones pueden estar fuera del intervalo { 0 ,1 } lo cual puede hacer difícil la interpretación como probabilidades.

- Para analizar la respuesta como probabilidades debemos usar logistic regression.

Tipos de classifieras

Non-linear classifiers:

- **K-Nearest Neighbors**
- Support Vector Machines
- Decision trees
- Boosted trees
- Random Forest
- Neural Networks

Linear classifiers

- Logistic regression
- **Naive Bayes Classifier**

Tipos de classifiers

Step 1

Divide dataset into a training set and a test set.

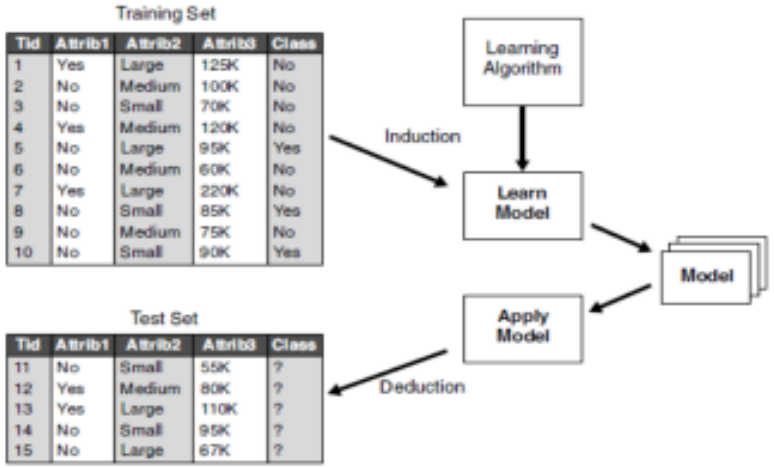
A **training set** consists of records whose **class labels are known**.

A **test set** contains records with **unknown class labels** (unknown for the algorithm).

Step 2

The **training set** is used **to build** a classification **model**.

The classification **model is applied to the test set**.



Step 3

Evaluate the performance of the classification model using the **confusion matrix**.

Confusion matrix is the tabulated counts of correctly and incorrectly predicted test records.

A confusion matrix can be **summarised in a performance metric**, e.g. accuracy or error rate.

Evaluamos el modelo

Confusion Matrix

		Predicted Class	
		Class = 1	Class = 0
Actual Class	Class = 1	f_{11}	f_{10}
	Class = 0	f_{01}	f_{00}

$$\textbf{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} = \frac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{01} + f_{00}}$$

$$\textbf{Error Rate} = \frac{\text{Number of incorrect predictions}}{\text{Total number of predictions}} = \frac{f_{10} + f_{01}}{f_{11} + f_{10} + f_{01} + f_{00}}$$

Most classification algorithms **seek models** that attain the **highest accuracy**, or equivalently, the **lowest error rate** when applied to the test set.

Agenda

1. Logistic regression

INFOX
educación sin fronteras

1. El modelo logístico

- Consideremos de nuevo la base de datos del banco, y la respuesta de default cae en dos categorías SI o NO.
- El Modelo logístico modela la probabilidad que Y pertenezca a una categoría en particular.
- Para nuestra base de datos de ejemplo, el modelo logístico modela el default.

$$\Pr(\texttt{default} = \texttt{Yes}|\texttt{balance}).$$

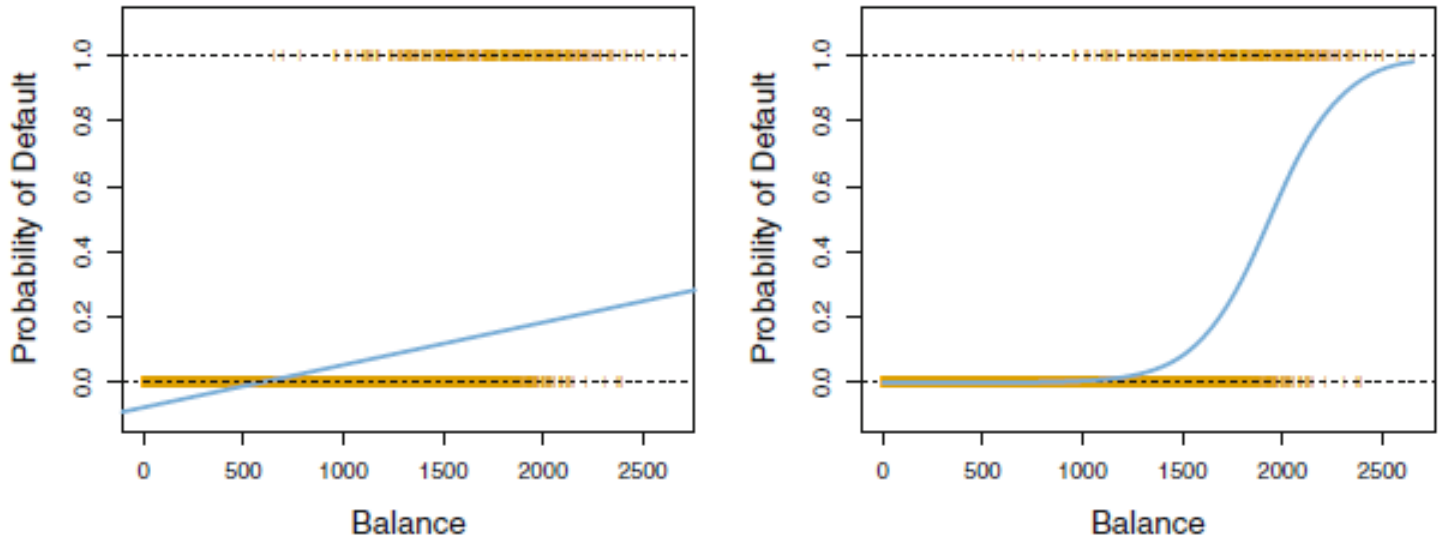
- Los valores de $\Pr(\texttt{default} = \texttt{Yes}|\texttt{balance})$., que abreviamos como $p(\texttt{balance})$ van en un rango entre 0 y 1.
- Luego, para cualquier valor del balance se puede hacer una predicción. Por ejemplo, podemos predecir Default = SI para cualquier individuo cuyo $p(\texttt{balance})$ > 0.5
- Alternativamente, si una empresa buscar ser mas conservativa en predecir individuos que ya están en riesgo de default, pueden usar un $p(\texttt{balance})$ > 0.1

1. El modelo logístico

- Cómo debemos modelar la relación entre $p(X) = \Pr(Y = 1|X)$ y X?
- En la sección anterior conversamos sobre un modelo de regresión lineal para representar esas probabilidades:

$$p(X) = \beta_0 + \beta_1 X.$$

- Si usamos esta aproximación para predecir default = Si usando el balance como única variable, entonces obtenemos el modelo mostrado en la siguiente figura.
-



1. El modelo logístico

- El problema que vemos en el gráfico izquierdo es común cuando usamos un modelo de regresión lineal.

- Para evitar este problema debemos modelar p(x) como una función que nos da valores entre 0 y 1 para cualquier valor de X. En la regresión logística usamos la función logística:

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}.$$

- Para resolver este modelo, usamos máximum likelihood que discutiremos luego. La función logística siempre producirá una curva en forma de S y respecto al valor de X obtendremos una sensible predicción. Asimismo, en el grafico anterior vemos que este modelo captura mejor el comportamiento del default de los clientes.

- Con un poco de manipulación de la primera función, tenemos:

$$\frac{p(X)}{1 - p(X)} = e^{\beta_0+\beta_1 X}.$$

1. El modelo logístico

- La cantidad $p(X)/[1-p(X)]$ se llama the odds, y puede tomar cualquier valor entre 0 e infinito.
- Valores del odds cerca a 0 o infinito indica un pequeña o una alta probabilidad en el default, respectivamente.
- Por ejemplo, en promedio 1 de 5 personas con un odds de ¼ va a hacer $p(X) = 0.2$ que implica un odds de $\frac{0.2}{1-0.2} = 1/4$.
- Otro ejemplo, en promedio 9 de 10 personas harán default con un odds de 9, dado que p(x)= 0.9, lo que da un odds de $\frac{0.9}{1-0.9} = 9$
- Si tomamos logaritmo a ambos lados de la ecuación tenemos: (la ecuación de la izquierda se llamada logit.

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X.$$

1. El modelo logístico

- En un modelo de regresión clásicc β_1 nos da el promedio de cambio en Y asociado a una unidad de incremento en X.

- En un modelo de regresión logística, incrementar X en una unidad cambia los log odc β_1 or o equivalentemente multiplica los odds por e^{β_1} . Sin embargo, dado que la relación entre p(X) y X no es un línea derecha, β_1 no corresponde a un cambio en X.

- La cantidad que p(x) va a cambiar va a depender del actual valor de X.

- **Estimando los coeficientes:**

- Los coeficientes β_0 and β_1 son desconocidos, y deben ser estimados usando el training Dataset. Podríamos usar un MCO para realizar esta estimación. Sin embargo, también podríamos usar métodos no lineales como máximum likelihood.

1. El modelo logístico

- La intuición básica para estimar un máximum likelihood para un modelo de regresión logística es la siguiente:
- Buscamos estimar β_0 y β_1 tal que la probabilidad $\hat{p}(x_i)$ de default de cada individuo sea muy cercana a la observada en la data real.

- Se busca estimar la siguiente ecuación:

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'}))$$

- Los estimados $\hat{\beta}_0$ and $\hat{\beta}_1$ son elegidos para maximizar la función de likelihood. Esta es una aproximación bastante general que nos permite estimar modelos no lineales.

- La estimación de esta ecuación la haremos mediante el software R.

1. El modelo logístico

- En la siguiente tabla mostramos las estimaciones que son resultado de estimar la probabilidad de default usando la base de datos del banco.

- Vemos que $\hat{\beta}_1$ = 0.0055, esto indica que un incremento en el balance es asociado con un incremento en la probabilidad de default. Para ser precisos, una unidad adicional en el balance es asociado con un incremento del log odds del default en 0.0055 unidades.

	Coefficient	Std. error	Z-statistic	P-value
Intercept	−10.6513	0.3612	−29.5	<0.0001
balance	0.0055	0.0002	24.9	<0.0001

1. El modelo logístico

- La estimación del modelo ha sido mostrada en la tabla anterior y tiene muchas similitudes con el modelo de regresión lineal. Por ejemplo, podemos medir la precisión de los coeficientes estimados computando sus errores estándares.

- Además el p value tiene la misma connotación que en el modelo de regresión lineal.

- El intercepto es típicamente no de nuestro interés, ya que su objetivo principal es ajustar el promedio total de la variable dependiente.

1. El modelo logístico – haciendo predicciones

- Una vez que los coeficientes están estimados, es simple calcular la probabilidad de default para cualquier balance de tarjeta de crédito. Por ejemplo, usando los coeficientes estimados dados en la tabla anterior, predecimos la probabilidad de default para un individuo con un balance de $ 1 000.

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1{,}000}}{1+e^{-10.6513+0.0055\times 1{,}000}} = 0.00576,$$

- El cual es menor de 1%. De otra manera, la predicción de probabilidad de default para un individuo con un balance de $2 000 es mucho más alta (0.586).

1. El modelo logístico

- Hasta ahora hemos estimado el modelo solo tomando en cuenta una variable, ahora vamos a extender este análisis usando múltiples predictores.

- Por analogía, la ecuación sería la siguiente:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- Donde $X = (X_1, \ldots, X_p)$ son p predictores. Entonces la ecuación anterior puede ser escrita como:

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}.$$

- Así como en el modelo anterior, usamos el método de máximum likelihood para estimar los betas.
-

1. El modelo logístico

- En la siguiente tabla vemos la estimación del modelo de regresión logístico usando varios predictores

	Coefficient	Std. error	Z-statistic	P-value
Intercept	−10.8690	0.4923	−22.08	<0.0001
balance	0.0057	0.0002	24.74	<0.0001
income	0.0030	0.0082	0.37	0.7115
student[Yes]	−0.6468	0.2362	−2.74	0.0062

- Las variables balance y Student tienen un p value menor a 0.05 lo cual indica que son significativas. El coeficiente de la variable dummy student es negativo, lo cual indica que los estudiantes son menos propensos a hacer default que los no estudiantes.

1. El modelo logístico

- Sustituyendo los estimados en la regresión. Por ejemplo, un estudiante con un balance de $ 1 500 y un ingreso de $ 40 000 tiene una probabilidad de default de

$$\hat{p}(X) = \frac{e^{-10.869+0.00574\times 1{,}500+0.003\times 40-0.6468\times 1}}{1+e^{-10.869+0.00574\times 1{,}500+0.003\times 40-0.6468\times 1}} = 0.058.$$

- Y un no estudiante

$$\hat{p}(X) = \frac{e^{-10.869+0.00574\times 1{,}500+0.003\times 40-0.6468\times 0}}{1+e^{-10.869+0.00574\times 1{,}500+0.003\times 40-0.6468\times 0}} = 0.105.$$

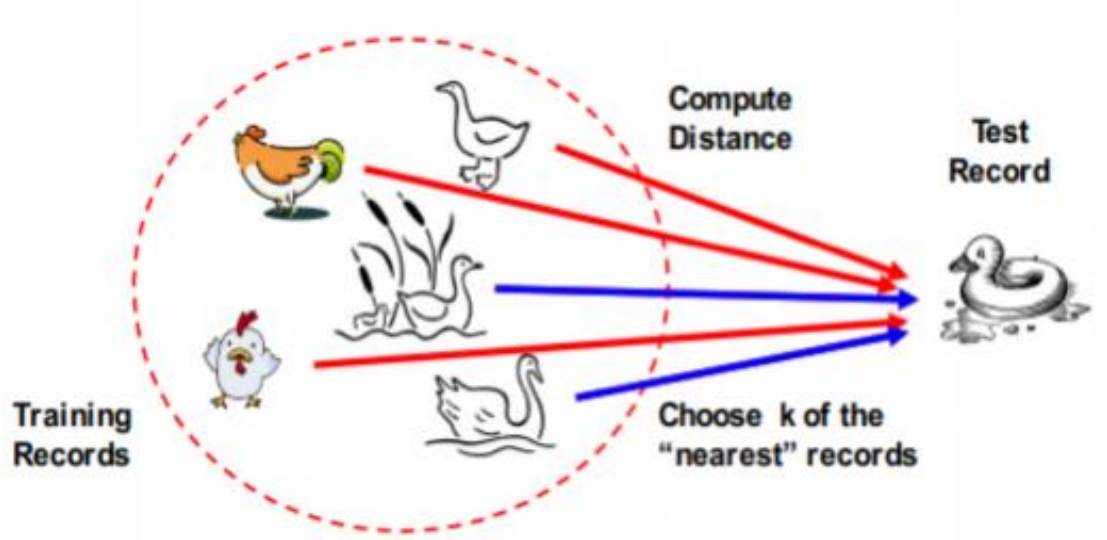
Agenda

2. K nearest neighbors

2. KNN

The K-NN algorithm classifies objects according to its nearest neighbours.

"If it walks like a duck, quacks like a duck, and looks like a duck, then it's probably a duck."



2. KNN

- **Non-linear classifier**
- **Lazy learner**
- **Non-parametric approach:** it does not make any assumption on data distribution (the data does not have to be normally distributed)

- **Instance-based learning:** uses specific training instances to make predictions without having to maintain an abstraction (or model) derived from data

- Makes predictions **based on local information -** which is derived through selecting *k*

- Can produce **arbitrarily shaped decision boundaries**: depends on k - can be more flexible & have higher variability

- Data pre-processing steps (**standardisation**) and decide on a **proximity measure** (typically Euclidean)

2. KNN

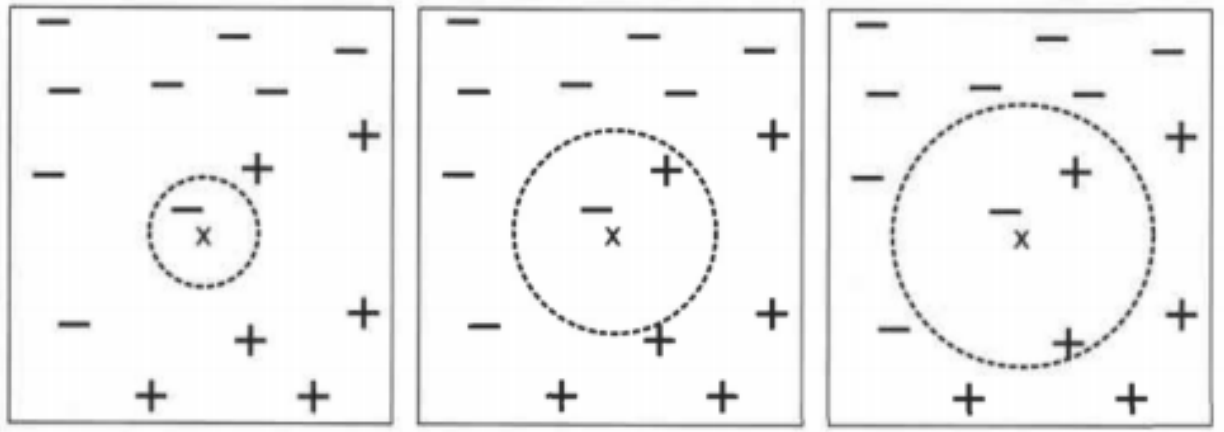
- Each example is a data point in a **d-dimensional space,** where ***d* = number of attributes**
- Given a test example, z, **compute its proximity to the rest of the data points** in the training set
- The **k-nearest neighbours of z refer to the k points closest to z**

1. x = data point we want to classify

1. On *Picture (a)*, x has one "**-**" neighbor;
   Thus x is assigned to the negative class.

1. On *Picture (b)*, x has two neighbors: a tie between "**-**" and "**+**" classes.
   Randomly choose one class to classify x.

1. On *Picture (c)*, there are two "**+**"s and one "**-**".
   Using majority voting scheme,
   x is assigned to the positive class.



(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

INFOX
educación sin fronteras

2. KNN - Algoritmo

1: Let k be the number of nearest neighbors and D be the set of training examples.
2: **for** each test example $z = (\mathbf{x}', y')$ **do**
3: Compute $d(\mathbf{x}', \mathbf{x})$, the distance between z and every example, $(\mathbf{x}, y) \in D$.
4: Select $D_z \subseteq D$, the set of k closest training examples to z.
5: $y' = \underset{v}{\operatorname{argmax}} \sum_{(\mathbf{x}_i, y_i) \in D_z} I(v = y_i)$
6: **end for**

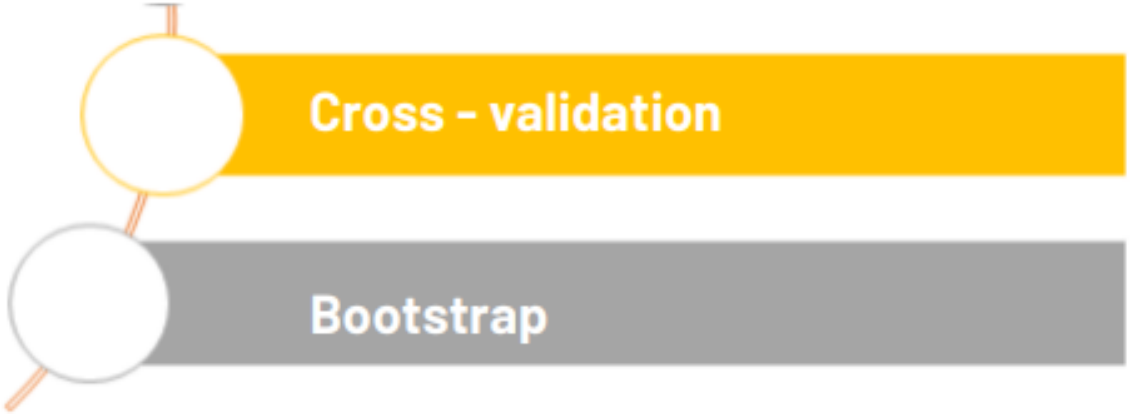
$$\text{Majority Voting: } y' = \underset{v}{\operatorname{argmax}} \sum_{(\mathbf{x}_i, y_i) \in D_z} I(v = y_i),$$

- In the majority voting approach, every neighbor has the same impact on the classification.
- This makes the algorithm sensitive to the choice of k.
- To solve this problem we can use distance-weighted voting, which assigns a weaker impact to the neighbors that are further away.

$$\text{Distance-Weighted Voting: } y' = \underset{v}{\operatorname{argmax}} \sum_{(\mathbf{x}_i, y_i) \in D_z} w_i \times I(v = y_i).$$

2. KNN – Escoger el óptimo K

- It is important to choose the right value for K (the correct level of flexibility)
- If K is too small, then classifier may be susceptible to overfitting (noise in training data)
- If K is too large, the classifier may misclassify the test instance (include far away data points)



With both methods we intend to choose the k that minimizes the average **error rate**:

$$\frac{\text{Number of incorrect predictions}}{\text{Total number of predictions}} = \frac{f_{10} + f_{01}}{f_{11} + f_{10} + f_{01} + f_{00}}$$

2. KNN – Escoger el óptimo K

Cross - validation

Example: K- folds cross validation

- Divides the sample into k groups (or folds)
- 1rst fold treated as validation set
- Model is fitted into the remaining k-1

Steps for choosing K

- Perform CV for different k values
- Compare the average error rate (number of misclassified observations)
- Choose the k which has the lowest average error rate

Application: Cross Validation with 10 folds

k	Toyota	Nissan
	Mean Absolute Error (Rs)	
1	45, 189	27, 258
3	54, 584	27, 134
5	62, 670	25, 741
10	63, 837	29, 289

2. KNN

Pros

- ★ Simple to implement
- ★ Works well with non-linear data
- ★ Flexible to feature/distance choices
- ★ Naturally handles multi-class cases
- ★ Can do well in practice with enough representative data

Cons

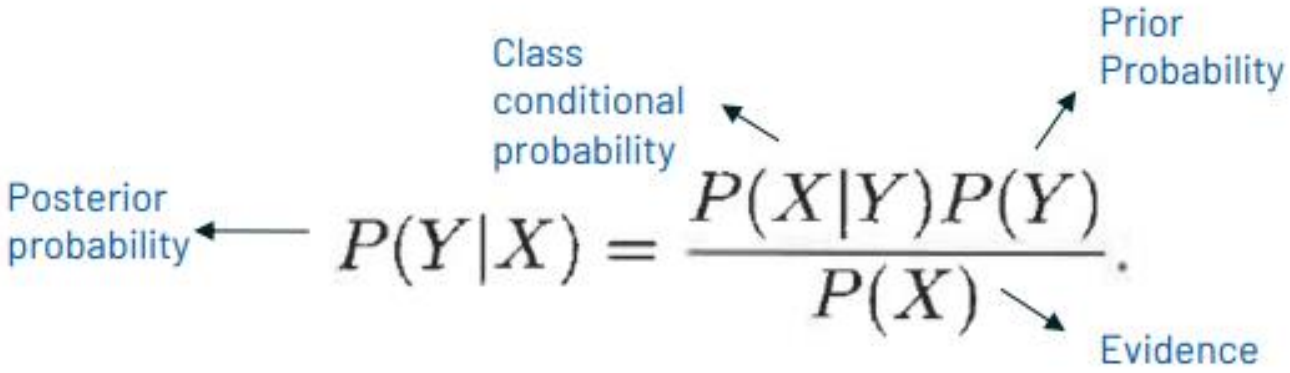
- ➢ How to determine value of K-NN
- ➢ High computation costs
- ➢ Must have a meaningful distance function
- ➢ Difficult to use in real time (lazy learner)

Agenda

3. Naive Bayes

3. Naive Bayes - ¿ Qué es?

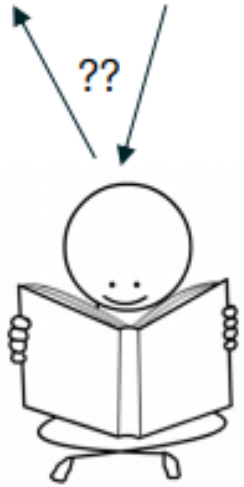
- **Linear classification method**
- **Eager learner**
- **Parametric:** summarizes data with a set of parameters of fixed size (independent of the size of the sample)
- **Probabilistic classifier** inspired by the Bayes theorem:



- **It assumes**:
  - that the attributes are conditionally independent, given the class label y

    (in order to estimate the class-conditional probability)

3. Naive Bayes - ¿ A qué nos referimos con condicionalmente independiente?



People with longer arms tend to have higher levels of reading skills.

This relationship can be explained by the factor, age.

A 6 years old child = short arms, lacks adult's reading skills

If age is fixed, then the observed relationship disappears.

Thus, arm length and reading skills are **conditionally independent** when age is fixed.

3. Naive Bayes - Ejemplo

A fruit may be an *apple* *if* it is:

- red
- round
- 3 inches in diameter

Even if these features depend on each other / other features, each property **independently contributes** to the **probability** that this fruit is an apple.

This is why it is known as Naïve.

3. Naive Bayes – Cómo funciona?

Step 1

With the **conditional independence assumption**, we estimate the conditional probability of each X_i given Y ($P(X|Y)$) in the training set.

We estimate the prior probability ($P(Y)$) as the proportion of Y class in the training set.

Step 2

To **classify a test record**, the classifier **computes the posterior probability,** $P(Y|X)$ for each class Y.

Step 3

Since $P(X)$ is fixed for every Y, it is sufficient to **choose the class that maximizes the numerator term.**

$$P(Y|\mathbf{X}) = \frac{P(Y)\prod_{i=1}^{d} P(X_i|Y)}{P(\mathbf{X})}$$

If $P(Yes|X) > P(No|X)$, then the record is classified as "Yes"

3. Naive Bayes – Estimando probabilidades condicionales

For a categorical attribute X_i, the conditional probability $P(X_i = x_i \mid Y = y)$ is estimated according to the fraction of training instances in class y that take on a particular attribute value x_i.

	binary	categorical	continuous	class
Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Example 1:

3 out of 7 people who repaid their loans also owned a home:

Conditional probability for $P(\text{Home Owner} = \text{Yes} \mid \text{No}) = 3/7$

Example 2:

Defaulted borrowers who are single:

Conditional probability for $P(\text{Marital Status} = \text{Single} \mid \text{Yes}) = 2/3$

3. Naive Bayes – Ejemplo

`head(fruitdata)`

Fruit	Long (x1)	Sweet (x2)	Yellow (x3)
Orange	0	1	0
Banana	1	0	1
Banana	1	1	1
Other	1	1	0
...	...	...	...

`Aggregated fruitdata`

Type	Long	Not Long	Sweet	Not Sweet	Yellow	Not Yellow	Total
Banana	400	100	350	150	450	50	500
Orange	0	300	150	150	300	0	300
Other	100	100	150	50	50	150	200
Total	500	500	650	350	800	200	1000

$$P(Y|X)$$

$$P(\text{Banana} \mid \textit{Long, Sweet and Yellow}) = \frac{P(X|Y)}{P(\text{Long}|\text{Banana}) * P(\text{Sweet}|\text{Banana}) * P(\text{Yellow}|\text{Banana})} * \frac{P(Y)}{P(\text{banana})} \Big/ \underbrace{P(\text{Long}) * P(\text{Sweet}) * P(\text{Yellow})}_{P(X)}$$

$$= \frac{0.8 * 0.7 * 0.9 * 0.5}{P(\text{Evidence})} = 0.252/P(\text{Evidence})$$

P(Orange | Long, Sweet and Yellow) = 0, because P(Long|Orange) = 0

P(Other Fruit | Long, Sweet and Yellow) = 0.01875/P(Evidence)

Answer: Banana - it has the highest probability amongst the 3 class

3. Naive Bayes – Para atributos continuos

There are two ways:

1. **Transform the continuous attributes into ordinal attributes** (through discretisation and replacement). The conditional probability is estimated by computing the fraction of training records belonging to class y that falls within the corresponding interval for X_i .

2. **Assume a certain form of probability distribution** for the continuous variable and estimate the parameters of the distribution using the training data. A **Gaussian** distribution is usually used to represent the class-conditional probability for continuous attributes.

¡GRACIAS!

INFOX
educación sin fronteras