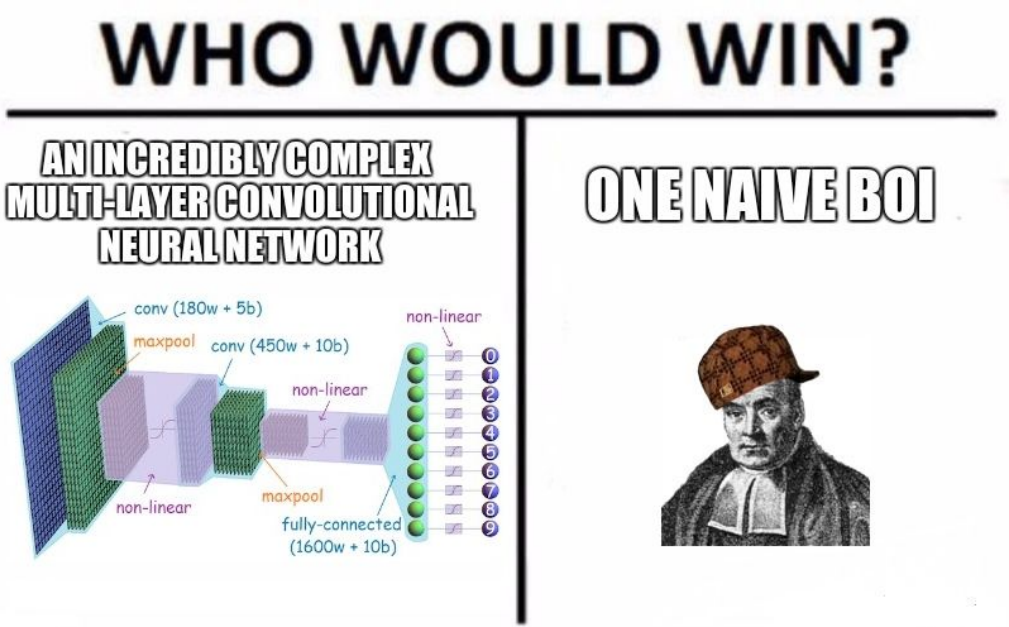


Laboratorio de Datos
Clase 21: Naive Bayes

El problema

Tengo una colección de documentos:

$X_1, X_2, X_3..., X_i, ..., X_n$

El problema

Tengo una colección de documentos:

$X_1, X_2, X_3..., X_i, ..., X_n$

Cada uno de ellos etiquetado:

$y_1, y_2, y_3, ..., y_i, ..., y_n \qquad y_i \in \{1, ..., k\}$

El problema

Tengo una colección de documentos:

$X_1, X_2, X_3..., X_i, ..., X_n$

Cada uno de ellos etiquetado:

$y_1, y_2, y_3, ..., y_i, ..., y_n$ $\quad$ $y_i \in \{1, ..., k\}$

Dada un nuevo documento, ¿cuánto valen estas probabilidades?

$p(y_i|X)$ $\quad$ i.e. probabilidad de tener la etiqueta i-ésima dado el documento

Un ejemplo de juguete

Clase 1 (CG):

X_1 = Tu amor me enseña a vivir, tu amor me enseña a sentir

X_2 = Estamos en la calle de la sensación, muy lejos del sol que quema de amor.

X_3 = La lágrima me dice, que yo tampoco soy, la hija de un amor, la hija del dolor

Un ejemplo de juguete

Clase 1 (CG):

X_1 = Tu amor me enseña a vivir, tu amor me enseña a sentir

X_2 = Estamos en la calle de la sensación, muy lejos del sol que quema de amor.

X_3 = La lágrima me dice, que yo tampoco soy, la hija de un amor, la hija del dolor

Clase 2 (2min):

X_4 = Ya no sos igual, ya no sos igual, sos un vigilante de la policia federal

X_5 = Y la policía entró, te marcó y te llevó, por haberlos puteado

X_6 = Barricada policial, hay que enfrentar, barricada policial, hay que transpasar

Un ejemplo de juguete

Clase 1 (CG):

X_1 = Tu amor me enseña a vivir, tu amor me enseña a sentir

X_2 = Estamos en la calle de la sensación, muy lejos del sol que quema de amor.

X_3 = La lágrima me dice, que yo tampoco soy, la hija de un amor, la hija del dolor

Clase 2 (2min):

X_4 = Ya no sos igual, ya no sos igual, sos un vigilante de la policia federal

X_5 = Y la policía entró, te marcó y te llevó, por haberlos puteado

X_6 = Barricada policial, hay que enfrentar, barricada policial, hay que transpasar

X = Y me enfrente al dolor, y cure mis heridas, y me encendí de amor ¿Clase 1 o 2?

Un ejemplo de juguete (procesado)

Clase 1 (CG):

X_1 = amor enseñar vivir amor enseñar sentir

X_2 = estar calle sensación lejos sol quemar amor

X_3 = lágrima decir ser hija amor hija dolor

Clase 2 (2min):

X_4 = ser igual ser igual vigilante policia federal

X_5 = policia entrar marcar llevar putear

X_6 = barricada policia enfrentar barricada policia transpasar

X = enfrentar dolor curar heridas encender amor

¿Clase 1 o 2?

$$P(A|B)P(B) = P(A,B)$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$p(CG|X), P(2min|X)$$

$$P(A|B)P(B) = P(A,B)$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$p(CG|X), P(2min|X)$$

$$P(2min|X) = \frac{P(X|2min)P(2min)}{P(X)}$$

$$P(A|B)P(B) = P(A,B)$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$p(CG|X), P(2min|X)$$

$$P(2min|X) = \frac{P(X|2min)P(2min)}{P(X)}$$

$$P(X|2min)P(2min) = P(X, 2min)$$

$$P(A|B)P(B) = P(A, B)$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$p(CG|X), P(2min|X)$$

$$P(2min|X) = \frac{P(X|2min)P(2min)}{P(X)}$$

$$P(A|B)P(B) = P(A,B)$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$P(X|2min)P(2min) = P(X,2min)$$

Esto no lo puedo estimar, porque el ejemplo no *X* no aparece en mis datos de entrenamiento

$$p(CG|X), P(2min|X)$$

$$P(2min|X) = \frac{P(X|2min)P(2min)}{P(X)}$$

$$P(A|B)P(B) = P(A,B)$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$P(X|2min)P(2min) = P(X,2min)$$

Esto no lo puedo estimar, porque el ejemplo no *X* no aparece en mis datos de entrenamiento

Pero... $X = (x_1, ..., x_d)$ es decir, *X* está hecha de palabras

$$p(CG|X), P(2min|X)$$

$$P(2min|X) = \frac{P(X|2min)P(2min)}{P(X)}$$

$$P(X|2min)P(2min) = P(X, 2min)$$

$$P(A|B)P(B) = P(A, B)$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Esto no lo puedo estimar, porque el ejemplo no *X* no aparece en mis datos de entrenamiento

Pero... $X = (x_1, ..., x_d)$ es decir, *X* está hecha de palabras

$$P(x_1, x_2, x_3, ..., x_d, 2min) = P(x_1|x_2, x_3, ..., x_d, 2min)P(x_2, x_3, ..., x_d, 2min)$$

$$p(CG|X), P(2min|X)$$

$$P(2min|X) = \frac{P(X|2min)P(2min)}{P(X)}$$

$$P(A|B)P(B) = P(A,B)$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$P(X|2min)P(2min) = P(X,2min)$$

Esto no lo puedo estimar, porque el ejemplo no *X* no aparece en mis datos de entrenamiento

Pero... $X = (x_1, ..., x_d)$ es decir, *X* está hecha de palabras

$$P(x_1, x_2, x_3, ..., x_d, 2min) = P(x_1|x_2, x_3, ..., x_d, 2min)P(x_2, x_3, ..., x_d, 2min)$$

$$P(x_1, x_2, x_3, ..., x_d, 2min) = P(x_1|x_2, x_3, ..., x_d, 2min)P(x_2|x_3, ..., x_d, 2min)P(x_3, ..., x_d, 2min)$$

$$p(CG|X), P(2min|X)$$

$$P(2min|X) = \frac{P(X|2min)P(2min)}{P(X)}$$

$$P(A|B)P(B) = P(A,B)$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$P(X|2min)P(2min) = P(X,2min)$$

Esto no lo puedo estimar, porque el ejemplo no *X* no aparece en mis datos de entrenamiento

Pero... $X = (x_1, ..., x_d)$ es decir, *X* está hecha de palabras

$$P(x_1, x_2, x_3, ..., x_d, 2min) = P(x_1|x_2, x_3, ..., x_d, 2min)P(x_2, x_3, ..., x_d, 2min)$$

$$P(x_1, x_2, x_3, ..., x_d, 2min) = P(x_1|x_2, x_3, ..., x_d, 2min)P(x_2|x_3, ..., x_d, 2min)P(x_3, ..., x_d, 2min)$$

$$P(x_1, x_2, x_3, ..., x_d, 2min) = P(x_1|x_2, x_3, ..., x_d, 2min)P(x_2|x_3, ..., x_d, 2min)...P(x_{d-1}|x_d, 2min)P(x_d|2min)P(2min)$$

$$p(CG|X), P(2min|X)$$

$$P(2min|X) = \frac{P(X|2min)P(2min)}{P(X)}$$

$$P(A|B)P(B) = P(A,B)$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$$P(X|2min)P(2min) = P(X,2min)$$

Esto no lo puedo estimar, porque el ejemplo no *X* no aparece en mis datos de entrenamiento

Pero... $X = (x_1, ..., x_d)$ es decir, *X* está hecha de palabras

$$P(x_1, x_2, x_3, ..., x_d, 2min) = P(x_1|x_2, x_3, ..., x_d, 2min)P(x_2, x_3, ..., x_d, 2min)$$

$$P(x_1, x_2, x_3, ..., x_d, 2min) = P(x_1|x_2, x_3, ..., x_d, 2min)P(x_2|x_3, ..., x_d, 2min)P(x_3, ..., x_d, 2min)$$

$$P(x_1, x_2, x_3, ..., x_d, 2min) = P(x_1|x_2, x_3, ..., x_d, 2min)P(x_2|x_3, ..., x_d, 2min)...P(x_{d-1}|x_d, 2min)P(x_d|2min)P(2min)$$

Hipotesis naive

Asumimos que la probabilidad de que ocurra una palabra determinada no está influenciada por la ocurrencia de las demás, únicamente depende de la categoría del documento

$$P(x_1, x_2, x_3, ..., x_d, 2min) = P(x_1|x_2, x_3, ..., x_d, 2min)P(x_2|x_3, ..., x_d, 2min)...P(x_{d-1}|x_d, 2min)P(x_d|2min)P(2min)$$

Hipotesis naive

Asumimos que la probabilidad de que ocurra una palabra determinada no está influenciada por la ocurrencia de las demás, únicamente depende de la categoría del documento

$$P(x_1, x_2, x_3, ..., x_d, 2min) = P(x_1|x_2, x_3, ..., x_d, 2min)P(x_2|x_3, ..., x_d, 2min)...P(x_{d-1}|x_d, 2min)P(x_d|2min)P(2min)$$

$$P(x_1|x_2, x_3, ..., x_d, 2min) = P(x_1|2min)$$

$$P(x_2|x_3, ..., x_d, 2min) = P(x_2|2min)$$

$$P(x_{d-1}|x_d, 2min) = P(x_{d-1}|2min)$$

Hipotesis Naive

Hipotesis naive

Asumimos que la probabilidad de que ocurra una palabra determinada no está influenciada por la ocurrencia de las demás, únicamente depende de la categoría del documento

$$P(x_1, x_2, x_3, ..., x_d, 2min) = P(x_1|x_2, x_3, ..., x_d, 2min)P(x_2|x_3, ..., x_d, 2min)...P(x_{d-1}|x_d, 2min)P(x_d|2min)P(2min)$$

$$P(x_1|x_2, x_3, ..., x_d, 2min) = P(x_1|2min)$$

$$P(x_2|x_3, ..., x_d, 2min) = P(x_2|2min)$$

$$P(x_{d-1}|x_d, 2min) = P(x_{d-1}|2min)$$

Hipotesis Naive

Entonces, usando esto llegamos a que

$$P(2min|X) \propto P(2min)P(x_1|2min)P(x_2|2min)...P(x_d|2min)$$

$$P(CG|X) \propto P(CG)P(x_1|CG)P(x_2|CG)...P(x_d|CG)$$

Hipotesis naive

Asumimos que la probabilidad de que ocurra una palabra determinada no está influenciada por la ocurrencia de las demás, únicamente depende de la categoría del documento

$$P(x_1, x_2, x_3, ..., x_d, 2min) = P(x_1|x_2, x_3, ..., x_d, 2min)P(x_2|x_3, ..., x_d, 2min)...P(x_{d-1}|x_d, 2min)P(x_d|2min)P(2min)$$

$$P(x_1|x_2, x_3, ..., x_d, 2min) = P(x_1|2min)$$

$$P(x_2|x_3, ..., x_d, 2min) = P(x_2|2min)$$

$$P(x_{d-1}|x_d, 2min) = P(x_{d-1}|2min)$$

Hipotesis Naive

Entonces, usando esto llegamos a que

$$P(2min|X) \propto P(2min)P(x_1|2min)P(x_2|2min)...P(x_d|2min)$$

$$P(CG|X) \propto P(CG)P(x_1|CG)P(x_2|CG)...P(x_d|CG)$$

La constante de proporcionalidad es la misma para ambas, *1/P(X)*
Predigo la etiqueta como el máximo entre estas dos probabilidades, así que no importa (multiplica a ambas por igual)

Hipotesis naive

Asumimos que la probabilidad de que ocurra una palabra determinada no está influenciada por la ocurrencia de las demás, únicamente depende de la categoría del documento

$$P(x_1, x_2, x_3, ..., x_d, 2min) = P(x_1|x_2, x_3, ..., x_d, 2min)P(x_2|x_3, ..., x_d, 2min)...P(x_{d-1}|x_d, 2min)P(x_d|2min)P(2min)$$

$$P(x_1|x_2, x_3, ..., x_d, 2min) = P(x_1|2min)$$

$$P(x_2|x_3, ..., x_d, 2min) = P(x_2|2min)$$

$$P(x_{d-1}|x_d, 2min) = P(x_{d-1}|2min)$$

Hipotesis Naive

Entonces, usando esto llegamos a que

$$P(2min|X) \propto P(2min)P(x_1|2min)P(x_2|2min)...P(x_d|2min)$$

$$P(CG|X) \propto P(CG)P(x_1|CG)P(x_2|CG)...P(x_d|CG)$$

Listo!

La constante de proporcionalidad es la misma para ambas, *1/P(X)*
Predigo la etiqueta como el máximo entre estas dos probabilidades, así que no importa (multiplica a ambas por igual)

Hagamos las cuentas

X = enfrentar dolor curar heridas encender amor

x_1 x_2 x_3 x_4 x_4 x_5

Hagamos las cuentas

X = enfrentar dolor curar heridas encender amor

x_1 x_2 x_3 x_4 x_4 x_5

$$P(2min|X) \propto P(2min)P(x_1|2min)P(x_2|2min)P(x_3|2min)P(x_4|2min)P(x_5|2min)P(x_6|2min)$$

Hagamos las cuentas

X = enfrentar dolor curar heridas encender amor

x_1 x_2 x_3 x_4 x_4 x_5

$$P(2min|X) \propto P(2min)P(x_1|2min)P(x_2|2min)P(x_3|2min)P(x_4|2min)P(x_5|2min)P(x_6|2min)$$

1/2 1/18 0/18 0/18 0/18 0/18 0/18

Hagamos las cuentas

X = enfrentar dolor curar heridas encender amor

x_1 x_2 x_3 x_4 x_4 x_5

$$P(2min|X) \propto P(2min)P(x_1|2min)P(x_2|2min)P(x_3|2min)P(x_4|2min)P(x_5|2min)P(x_6|2min)$$

1/2 1/18 0/18 0/18 0/18 0/18 0/18

Esto es un problema: si aparecen palabras en *X* que no estaban en mi set de entrenamiento, la probabilidad me da automáticamente 0.

Hagamos las cuentas

X = enfrentar dolor curar heridas encender amor

x_1 x_2 x_3 x_4 x_4 x_5

$$P(2min|X) \propto P(2min)P(x_1|2min)P(x_2|2min)P(x_3|2min)P(x_4|2min)P(x_5|2min)P(x_6|2min)$$

1/2 1/18 0/18 0/18 0/18 0/18 0/18

Esto es un problema: si aparecen palabras en *X* que no estaban en mi set de entrenamiento, la probabilidad me da automáticamente 0.

Suavizado Laplaciano: $\frac{x_i}{N} \rightarrow \frac{x_i+\alpha}{N+\alpha K}$ donde α es un parámetro y *K* la cantidad de palabras distintas

Hagamos las cuentas

X = enfrentar dolor curar heridas encender amor

x_1 x_2 x_3 x_4 x_4 x_5

$$P(2min|X) \propto P(2min)P(x_1|2min)P(x_2|2min)P(x_3|2min)P(x_4|2min)P(x_5|2min)P(x_6|2min)$$ **= 1.378e-10**

1/2 ~~1/18~~ ~~0/18~~ ~~0/18~~ ~~0/18~~ ~~0/18~~ ~~0/18~~

$K = 26, \alpha = 1$ 2/44 1/44 1/44 1/44 1/44 1/44

Esto es un problema: si aparecen palabras en *X* que no estaban en mi set de entrenamiento, la probabilidad me da automáticamente 0.

Suavizado Laplaciano: $\frac{x_i}{N} \rightarrow \frac{x_i+\alpha}{N+\alpha K}$ donde α es un parámetro y *K* la cantidad de palabras distintas

Hagamos las cuentas

X = enfrentar dolor curar heridas encender amor

x_1 x_2 x_3 x_4 x_4 x_5

$$P(CG|X) \propto P(CG)P(x_1|CG)P(x_2|CG)P(x_3|CG)P(x_4|CG)P(x_5|CG)P(x_6|CG)$$

Hagamos las cuentas

X = enfrentar dolor curar heridas encender amor

x_1 x_2 x_3 x_4 x_4 x_5

$$P(CG|X) \propto P(CG)P(x_1|CG)P(x_2|CG)P(x_3|CG)P(x_4|CG)P(x_5|CG)P(x_6|CG)$$ **= 5.277e-10**

1/2 1/46 2/46 1/46 1/46 1/46 5/46

$$K = 26, \alpha = 1$$

Hagamos las cuentas

X = enfrentar dolor curar heridas encender amor

x_1 x_2 x_3 x_4 x_4 x_5

$$P(CG|X) \propto P(CG)P(x_1|CG)P(x_2|CG)P(x_3|CG)P(x_4|CG)P(x_5|CG)P(x_6|CG)$$ **= 5.277e-10**

1/2 1/46 2/46 1/46 1/46 1/46 5/46

$K = 26, \alpha = 1$

La predicción es la clase CG (la probabilidad da unas 4 veces más que para la clase 2min)

Naive Bayes con dos clases

$$p(D \mid S) = \prod_i p(w_i \mid S) \qquad p(D \mid \neg S) = \prod_i p(w_i \mid \neg S)$$

Naive Bayes con dos clases

$$p(D \mid S) = \prod_i p(w_i \mid S) \qquad p(D \mid \neg S) = \prod_i p(w_i \mid \neg S)$$

$$p(S \mid D) = \frac{p(S)}{p(D)} \prod_i p(w_i \mid S)$$

$$p(\neg S \mid D) = \frac{p(\neg S)}{p(D)} \prod_i p(w_i \mid \neg S)$$

Naive Bayes con dos clases

$$p(D \mid S) = \prod_i p(w_i \mid S) \qquad p(D \mid \neg S) = \prod_i p(w_i \mid \neg S)$$

$$p(S \mid D) = \frac{p(S)}{p(D)} \prod_i p(w_i \mid S)$$

$$p(\neg S \mid D) = \frac{p(\neg S)}{p(D)} \prod_i p(w_i \mid \neg S)$$

$$\frac{p(S \mid D)}{p(\neg S \mid D)} = \frac{p(S)}{p(\neg S)} \prod_i \frac{p(w_i \mid S)}{p(w_i \mid \neg S)}$$

Naive Bayes con dos clases

$$p(D \mid S) = \prod_i p(w_i \mid S) \qquad p(D \mid \neg S) = \prod_i p(w_i \mid \neg S)$$

$$p(S \mid D) = \frac{p(S)}{p(D)} \prod_i p(w_i \mid S)$$

$$p(\neg S \mid D) = \frac{p(\neg S)}{p(D)} \prod_i p(w_i \mid \neg S)$$

$$\frac{p(S \mid D)}{p(\neg S \mid D)} = \frac{p(S)}{p(\neg S)} \prod_i \frac{p(w_i \mid S)}{p(w_i \mid \neg S)}$$

tomo log

$$\ln \frac{p(S \mid D)}{p(\neg S \mid D)} = \ln \frac{p(S)}{p(\neg S)} + \sum_i \ln \frac{p(w_i \mid S)}{p(w_i \mid \neg S)}$$

Naive Bayes con dos clases

$$\ln \frac{p(S \mid D)}{p(\neg S \mid D)} = \ln \frac{p(S)}{p(\neg S)} + \sum_i \ln \frac{p(w_i \mid S)}{p(w_i \mid \neg S)}$$

Naive Bayes con dos clases

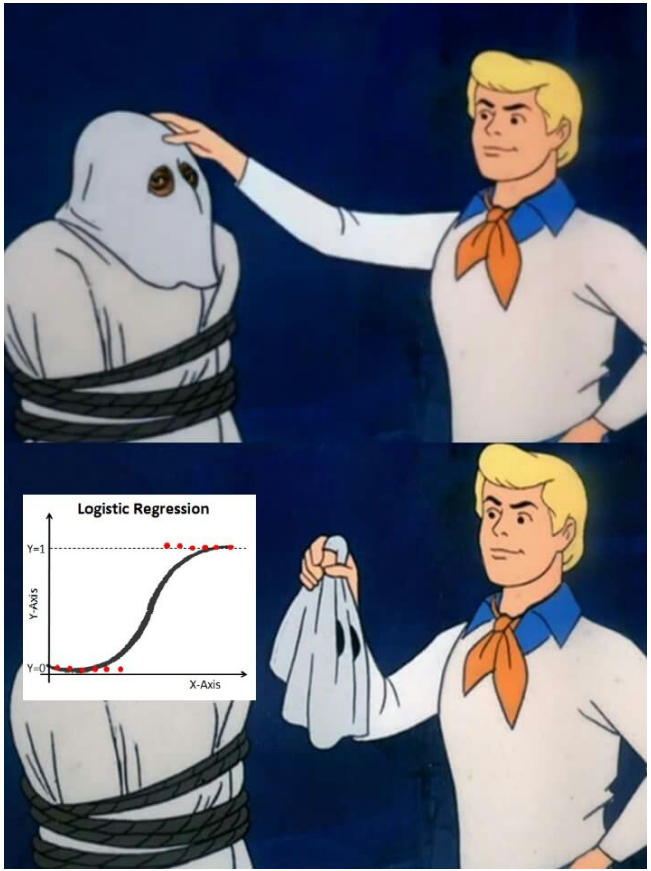
$$\ln \frac{p(S \mid D)}{p(\neg S \mid D)} = \ln \frac{p(S)}{p(\neg S)} + \sum_i \ln \frac{p(w_i \mid S)}{p(w_i \mid \neg S)}$$

$$\ln \frac{\mathrm{p}}{1-\mathrm{p}} = \beta_0 + \sum_i \beta_i x_i$$

Naive Bayes con dos clases

$$\ln \frac{p(S \mid D)}{p(\neg S \mid D)} = \ln \frac{p(S)}{p(\neg S)} + \sum_i \ln \frac{p(w_i \mid S)}{p(w_i \mid \neg S)}$$

$$\ln \frac{\mathrm{p}}{1-\mathrm{p}} = \beta_0 + \sum_i \beta_i x_i$$



sklearn.naive_bayes.MultinomialNB

class `sklearn.naive_bayes.MultinomialNB`(*, *alpha=1.0*, *fit_prior=True*, *class_prior=None*) [source]

Naive Bayes classifier for multinomial models

The multinomial Naive Bayes classifier is suitable for classification with discrete features (e.g., word counts for text classification). The multinomial distribution normally requires integer feature counts. However, in practice, fractional counts such as tf-idf may also work.

Read more in the User Guide.

Parameters:

alpha : *float, default=1.0*
Additive (Laplace/Lidstone) smoothing parameter (0 for no smoothing).

fit_prior : *bool, default=True*
Whether to learn class prior probabilities or not. If false, a uniform prior will be used.

class_prior : *array-like of shape (n_classes,), default=None*
Prior probabilities of the classes. If specified the priors are not adjusted according to the data.