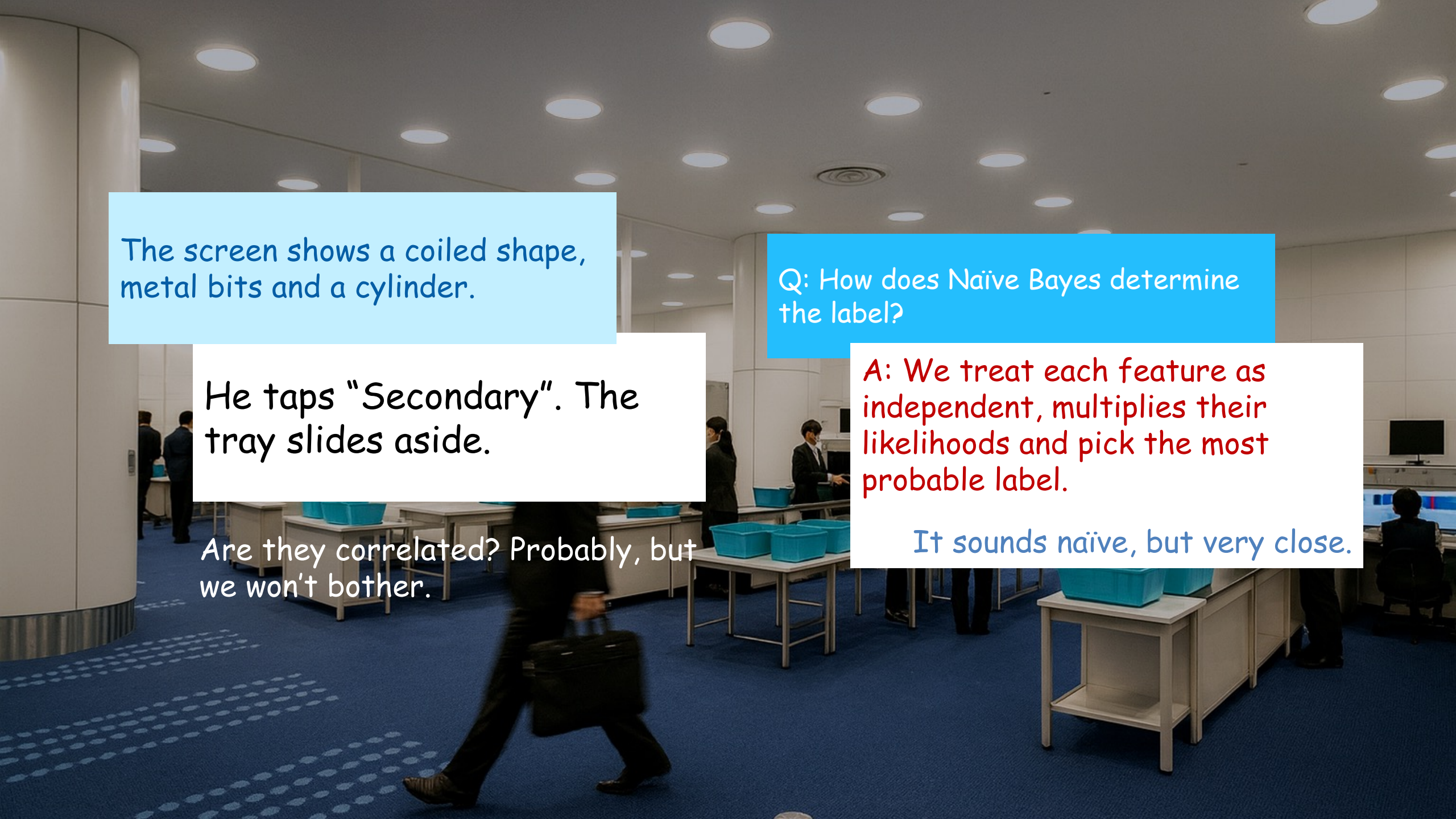
Machine Learning

Naïve Bayes

Tarapong Sreenuch

8 February 2024

克明峻德，格物致知



The screen shows a coiled shape, metal bits and a cylinder.

He taps "Secondary". The tray slides aside.

Are they correlated? Probably, but we won't bother.

Q: How does Naïve Bayes determine the label?

A: We treat each feature as independent, multiplies their likelihoods and pick the most probable label.

It sounds naïve, but very close.

2015 Gallup Poll: Online Dating Sites

		Age				
		18-29	30-49	50-64	65+	Total
Used online dating site	Yes	60	86	58	21	225
	No	255	426	450	382	1513
	Total	315	512	508	403	1738

% of 30-49 year olds using online dating sites = 86/512 ≈ 0.17.

Conditional Probability

Definition: The conditional probability of event A given an event B happened is

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

where we assume $P(B) \neq 0$).

An equivalent and useful formula is

$$P(A \cap B) = P(A|B)P(B)$$

Conditional Probability (cont.)

		Age				
		18-29	30-49	50-64	65+	Total
Used online dating site	Yes	60	86	58	21	225
	No	255	426	450	382	1513
	Total	315	512	508	403	1738

$$P(\text{Used}|30-49) = \frac{P(30-49 \cap \text{Used})}{P(30-49)}$$

$$= \frac{\frac{86}{1738}}{\frac{512}{1738}} = \frac{86}{512} \approx 0.17$$

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

Bayes Theorem

Theorem (Bayes Rule): For events A and B, where $P(A), P(B) > 0$,

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

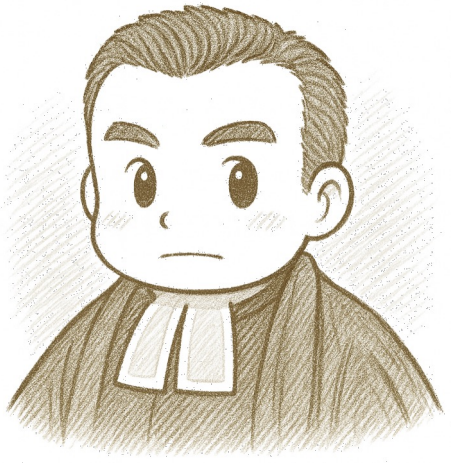
$$P(A \cap B) = P(B \cap A)$$

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

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

$P(A), P(B) > 0$, is called the prior (our belief without knowing anything).

$P(A|B)$ is called the posterior (our belief after learning B).



Bayes Theorem (cont.)

		Age				
		18-29	30-49	50-64	65+	Total
Used online dating site	Yes	60	86	58	21	225
	No	255	426	450	382	1513
	Total	315	512	508	403	1738

$$P(30{-}49|\text{Used}) = \frac{P\,(\text{Used}|30-49)\,P(30-49)}{P(\text{Used})}$$

$$= \frac{\frac{86}{512} \times \frac{512}{1738}}{\frac{225}{1738}} = \frac{86}{225} \approx 0.38$$

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

Recall: Joint Probability

Q: What is the probability (or likelihood) of rolling a six on both dice?

A: $\frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$

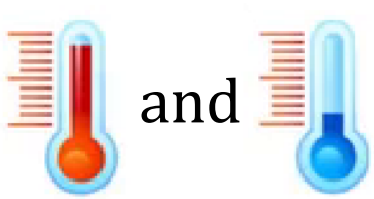
$$P(A,B) = P(A) \times P(B)$$

Probability of Rolling a Six on Dice A

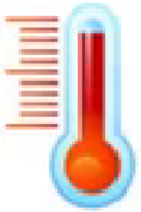
Probability of Rolling a Six on Dice B

Joint Probability: Counter Example

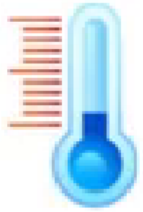
$$P(A, B) = P(A) \times P(B)$$


and

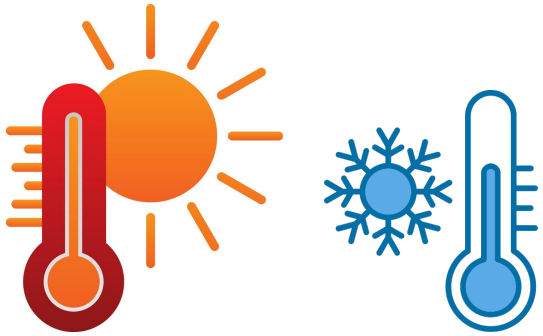
$= 0$



> 0



> 0



Events often exhibit correlations. It is unlikely to hold true in most real-world scenarios

Naïve Assumption

The naïve assumption refers to the simplifying assumption of conditional independence. It assumes that all the features (variables) in the dataset are independent of each other, given a target class label.

$$P(x_1, x_2, \ldots, x_M|Y) \approx P(x_1|Y)P(x_2|Y)\cdots P(x_M|Y)$$
$$\approx \prod_{i=1}^{M} P(x_i|Y)$$

Why Is It Called "Naïve"?

Features often exhibit correlations. It is unlikely to hold true in most real-world scenarios

Naïve Bayes Classifier

The Naïve Bayes classifier relies on Bayes' Theorem:

Conditional Probability (or Likelihood) *Prior Probability*

$$P(y \mid x_1, x_2, \ldots, x_D) = \frac{P(x_1, x_2, \ldots, x_D \mid y)P(y)}{P(x_1, x_2, \ldots, x_D)}$$

Posterior Probability *Evidence*

$$\approx \frac{P(x_1|y)P(x_2|y)\cdots P(x_D|y)P(y)}{P(x_1, x_2, \ldots, x_D)}$$

$$\propto P(x_1|y)P(x_2|y)\cdots P(x_D|y)P(y)$$

Bayes' Theorem: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

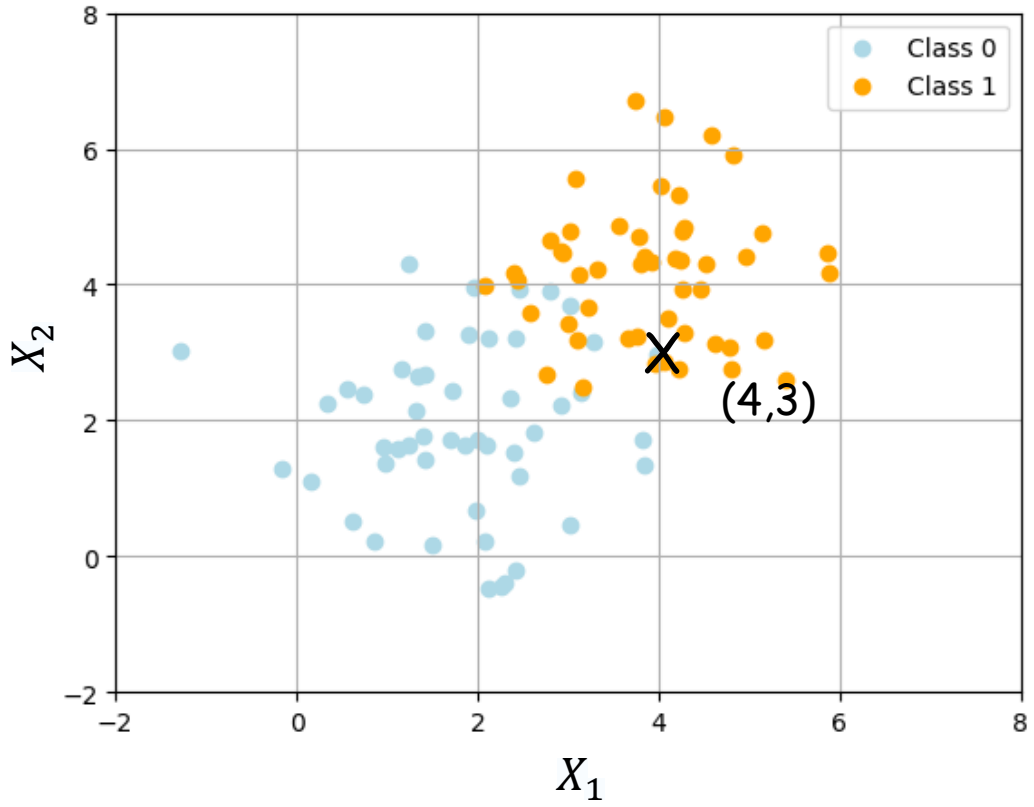
Naïve Assumption:

$$P(x_1, x_2, \ldots, x_D|y) \approx P(x_1|y)P(x_2|y)\cdots P(x_D|y)$$

Ignoring the Evidence:

$P(x_1, x_2, \ldots, x_D)$ *is constant for all classes.*

Ignoring Evidence



$$P(y = 0\ |\vec{x} = (4,3)) \propto \frac{P(x_1 = 4|y = 0)P(x_2 = 3|y = 0)}{P(\vec{x} = (4,3))}$$

$$P(y = 1\ |\vec{x} = (4,3)) \propto \frac{P(x_1 = 4|y = 1)P(x_1 = 3|y = 1)}{P(\vec{x} = (4,3))}$$

Evidence

Think About It: (Example)

0.7 > 0.3 → 0.7/x > 0.3/x regardless what the value of x (>0) will be.

To predict the class y, we selects the class based on which of the twos, i.e. $P\ (y = 0\ |\vec{x} = (4,3))$ and$P\ (y = 1\ |\vec{x} = (4,3))$, is the highest. Both are sharing the same evidence, i.e. $P(4,3)$, and hence we can drop it from the posterior calculations. This reduces the calculations to $P\ (y = 0\ |\vec{x} = (4,3)) \propto P(x_1 = 4|y = 0)P(x_2 = 3|y = 0)$ and $P\ (y = 1\ |\vec{x} = (4,3)) \propto P(x_1 = 4|y = 1)P(x_1 = 3|y = 1)$.

Naïve Bayes Classifier

The Naïve Bayes classifier relies on Bayes' Theorem:

Conditional Probability (or Likelihood) *Prior Probability*

$$P(y \mid x_1, x_2, \ldots, x_D) = \frac{P(x_1, x_2, \ldots, x_D \mid y)P(y)}{P(x_1, x_2, \ldots, x_D)}$$

Posterior Probability *Evidence*

$$\approx \frac{P(x_1|y)P(x_2|y)\cdots P(x_D|y)P(y)}{P(x_1, x_2, \ldots, x_D)}$$

$$\propto P(x_1|y)P(x_2|y)\cdots P(x_D|y)P(y)$$

$$\propto P(y)\prod_{i=1}^{D} P(x_i|y)$$

Bayes' Theorem: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

Naïve Assumption:

$P(x_1, x_2, \ldots, x_D|y) \approx P(x_1|y)P(x_2|y)\cdots P(x_D|y)$

Ignoring the Evidence:

$P(x_1, x_2, \ldots, x_D)$ is constant for all classes.

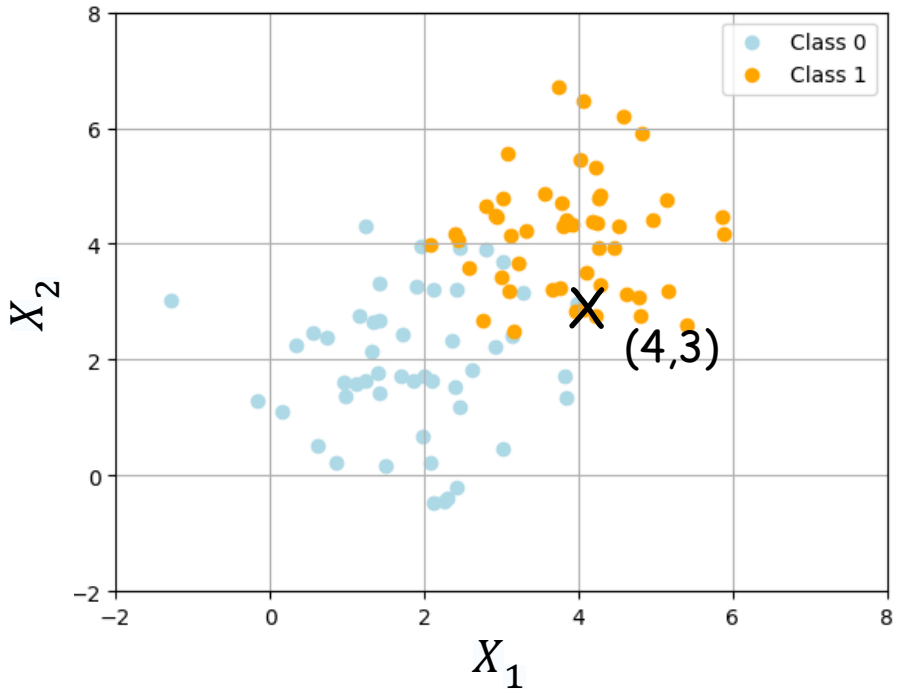
To predict the class y for a given input $\vec{x}$, the Naïve Bayes classifier selects the class with the highest posterior probability:

$$\hat{y} = \arg\max_y \left(P(y)\prod_{i=1}^{D} P(x_i|y)\right)$$

Gaussian Naïve Bayes

To predict the class $\mathbf{y}$, we calculate the likelihood of a data point $\vec{\mathbf{x}}$ being class $\mathbf{y}$, and then select the class with the highest posterior probability.

$$P(y|x_1, x_2) \propto P(x_1|y\ )P(x_2|y\ )P(y)$$



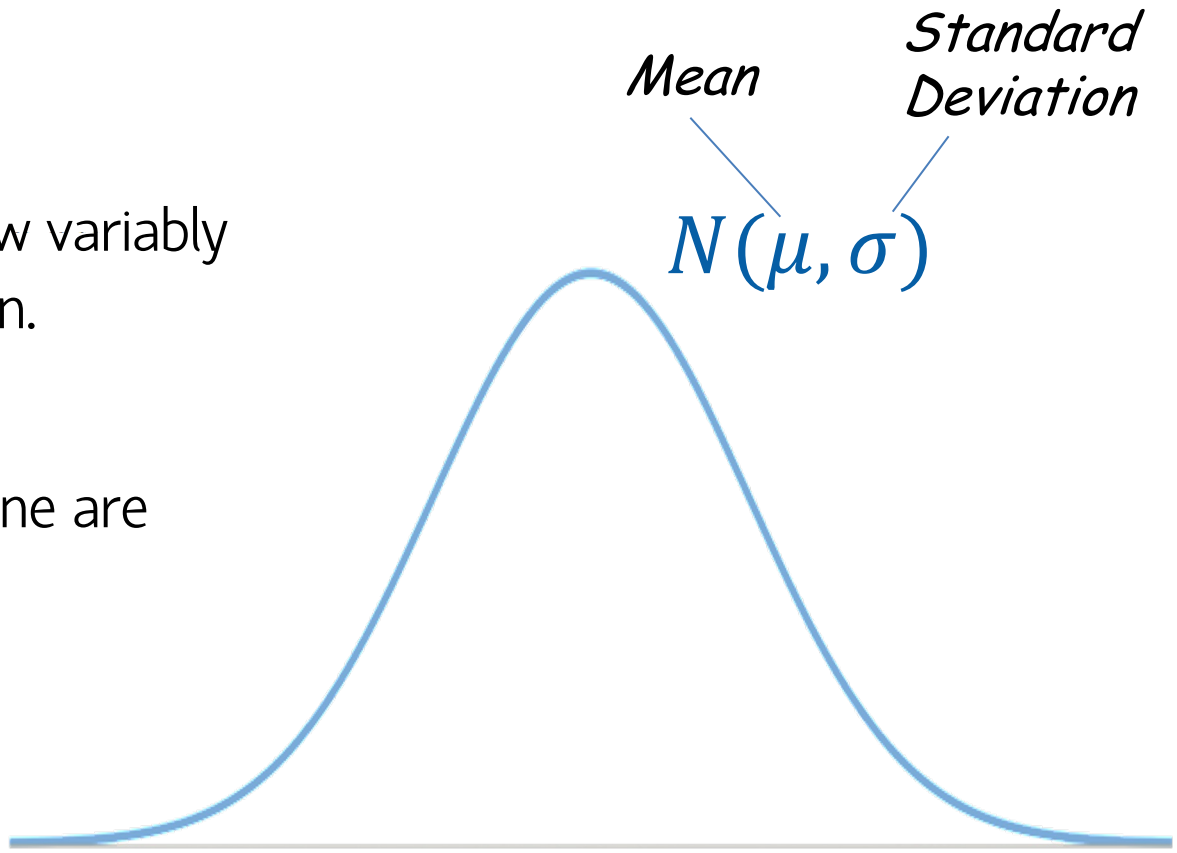
Q: How do we calculate these?

A: We count frequencies for prior probabilities. For conditional probabilities, we fit Normal distributions from the data and then compute the likelihood.

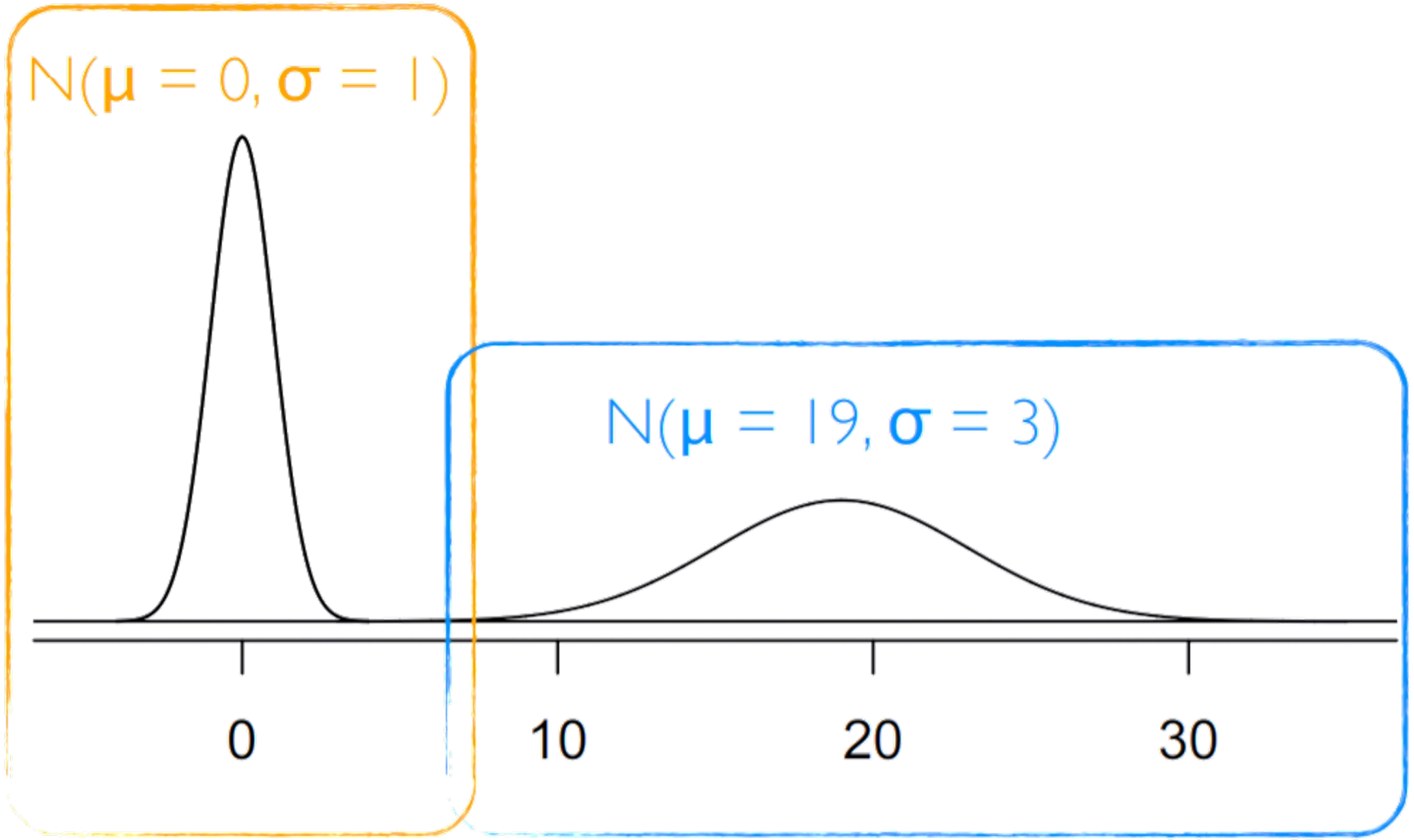
Normal Distribution

- Unimodal and Symmetric
  - Bell Curve
- It follows very strict guidelines about how variably the data are distributed around the mean.
- Many variables are nearly normal, but none are exactly normal.

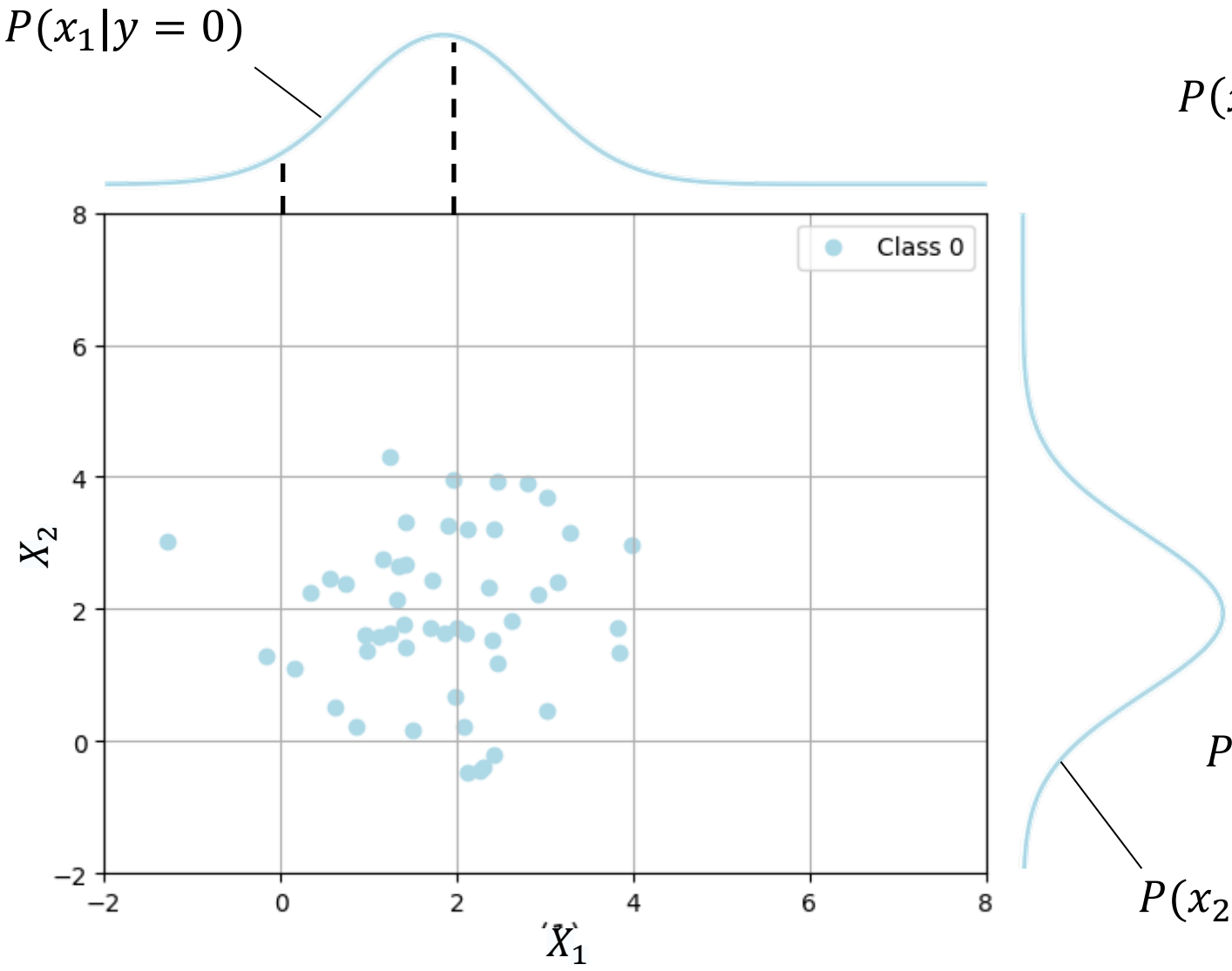
$$N(\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}$$



Normal Distribution (cont.)



Conditional Probability



$$P(x_1|y=0) = N(2,1.25)$$

$$= \frac{1}{\sqrt{2\pi \times 1.25^2}} e^{\left(-\frac{(x_1-2)^2}{2\times 1.25^2}\right)}$$

Examples: $P(x_1 = 2|y=0) = 0.319$

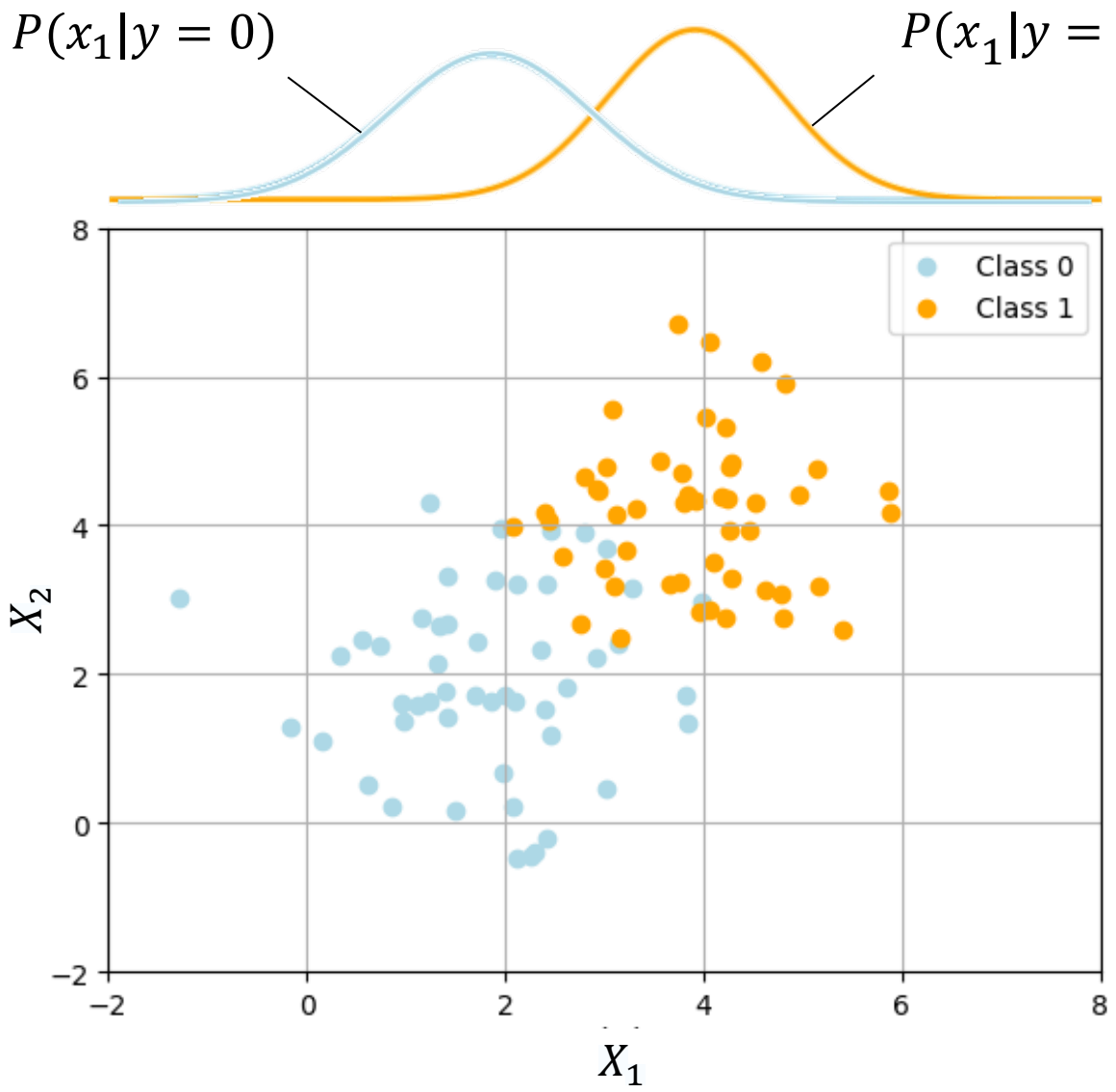
$P(x_1 = 0|y=0) = 0.089$

$$P(x_2|y=0) = N(2,1.25)$$

$$= \frac{1}{\sqrt{2\pi \times 1.25^2}} e^{(-\frac{(x_2-2)^2}{2\times 1.25^2})}$$

$P(x_2|y=0)$

Conditional Probability (cont.)

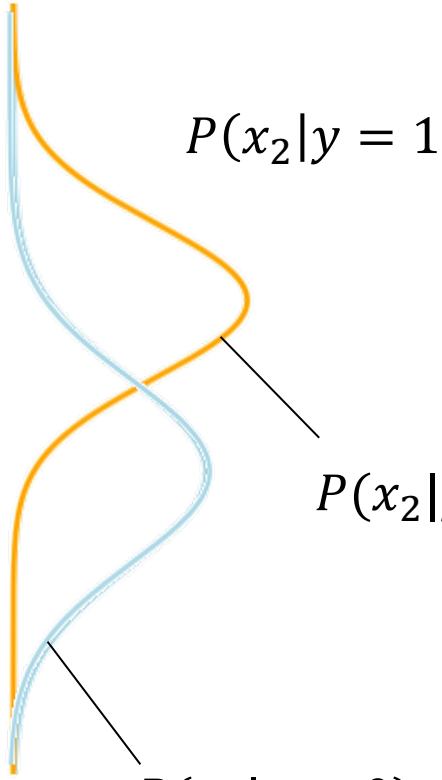


$$P(x_1|y=1) = N(4,1)$$

$$= \frac{1}{\sqrt{2\pi \times 1^2}} e^{\left(-\frac{(x_1-4)^2}{2\times 1^2}\right)}$$

$$P(x_2|y=1) = N(4,1)$$

$$= \frac{1}{\sqrt{2\pi \times 1^2}} e^{\left(-\frac{(x_2-4)^2}{2\times 1^2}\right)}$$



Posterior Probability

$$P(y = 1|x_1 = 4, x_2 = 3) = P(x_1 = 4|y = 0)P(x_2 = 3|y = 0)P(y = 0)$$

$$P(y = 0|x_1 = 4, x_2 = 3) = 0.087 \times 0.232 \times 0.5$$

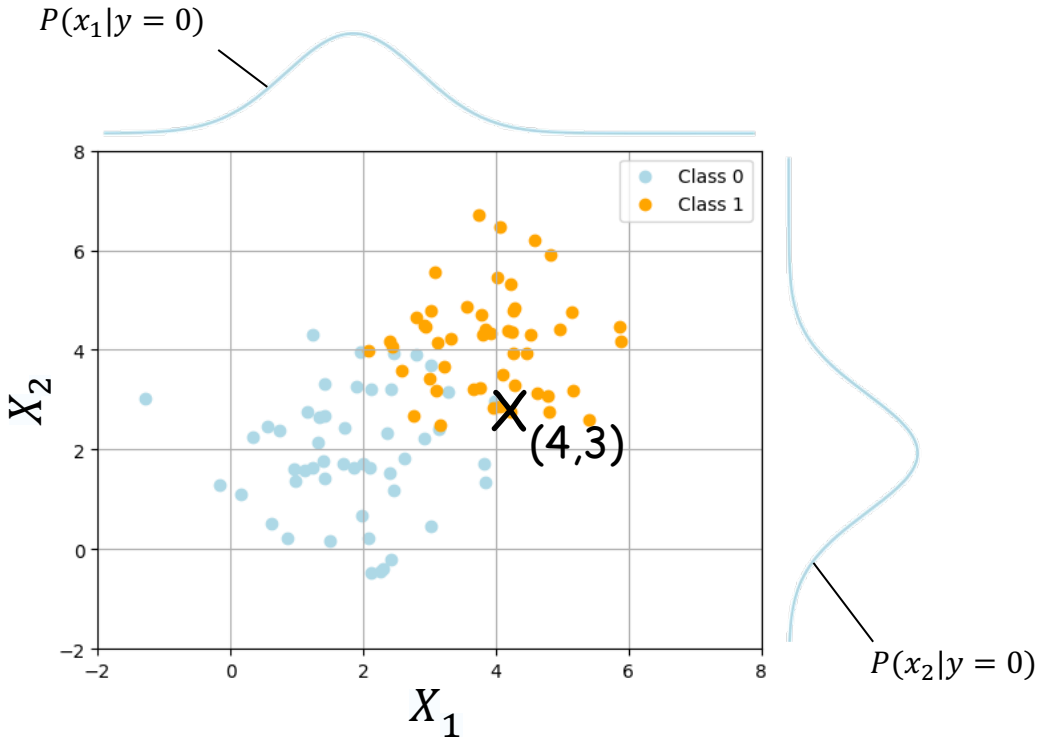
$$= 0.010$$

Prior Probability:

$$P(y = 0) = \frac{\text{Samples in Class 0}}{\text{Samples in Class 0} + \text{Samples in Class 1}}$$

$$= \frac{50}{50 + 50}$$

$$= 0.5$$



Posterior Probability (cont.)

$$P\,(y = 1|x_1 = 4, x_2 = 3) = P\,(x_1 = 4|y = 1)\,P\,(x_2 = 3|y = 1)\,P(y = 1)$$

$$P\,(y = 1|x_1 = 4, x_2 = 3) = 0.399 \times 0.242 \times 0.5$$

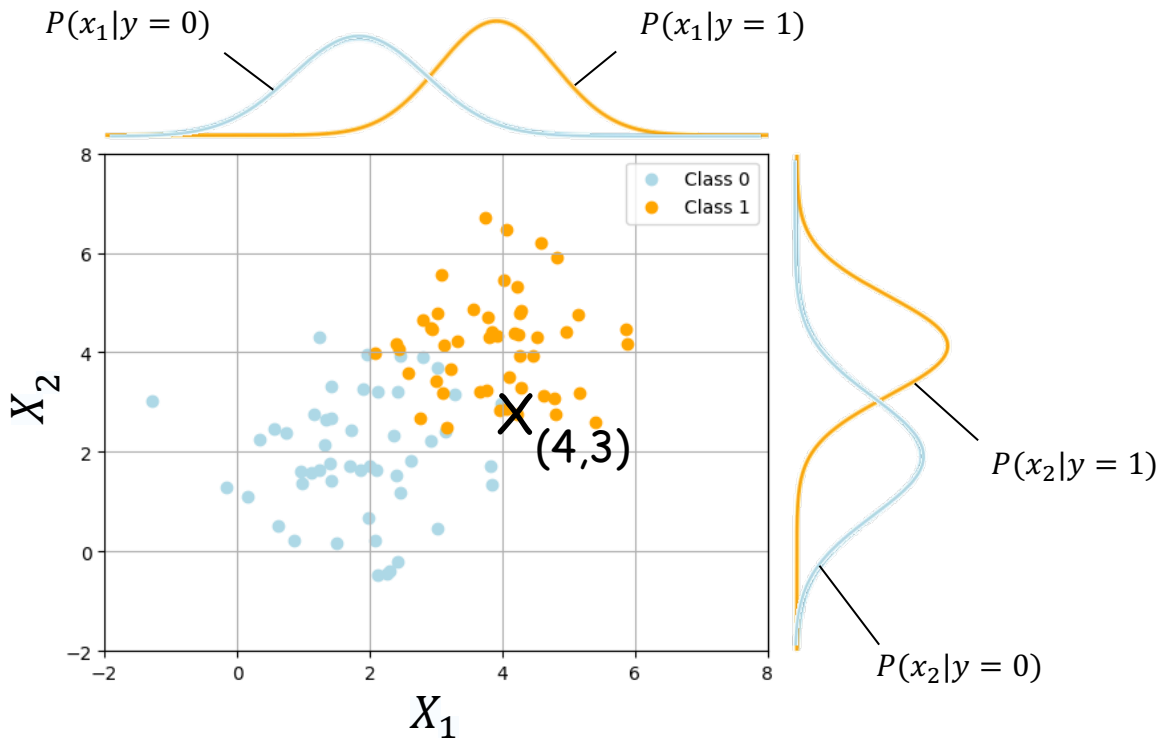
$$= 0.048$$

Prior Probability:

$$P(y = 1) = \frac{\text{Samples in Class 1}}{\text{Samples in Class 0} + \text{Samples in Class 1}}$$

$$= \frac{50}{50 + 50}$$

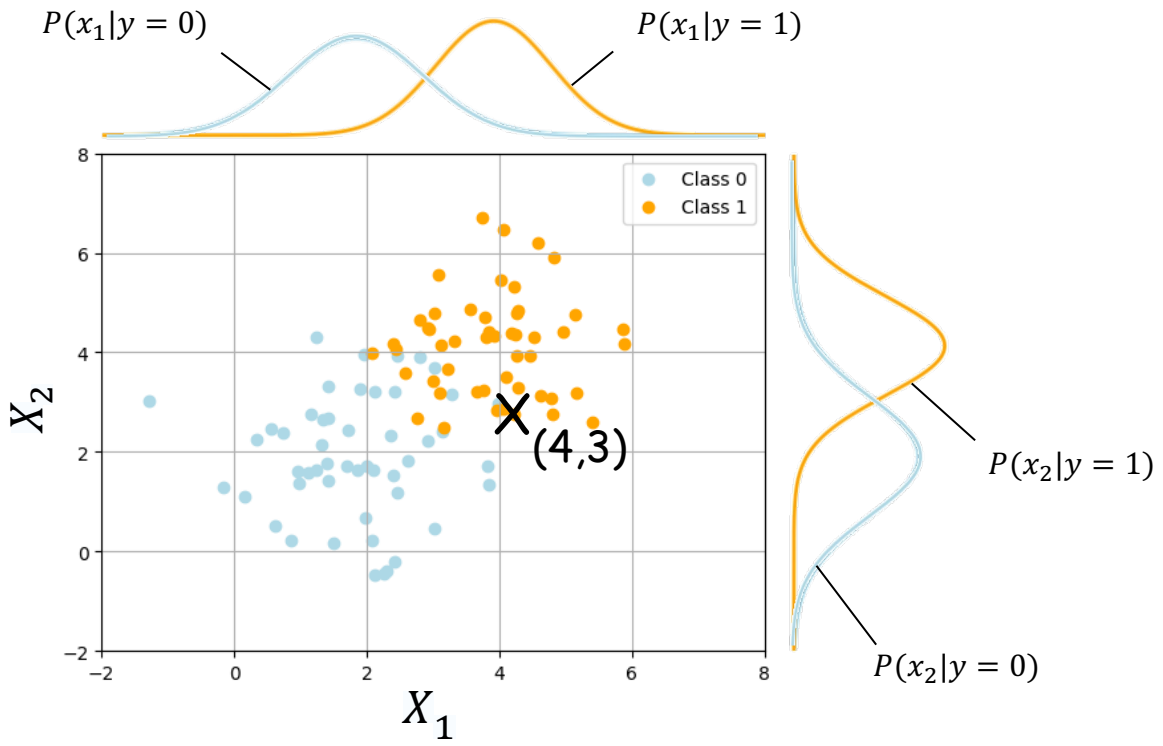
$$= 0.5$$



Prediction: Highest Posterior Probability

$$P\left(y = 0|\vec{x} = (4,3)\right) = 0.010 \qquad P\left(y = 1|\vec{x} = (4,3)\right) = 0.048$$

Since 0.048 > 0.010, (4,3) is predicted to belong to class 1.



Classification Rule

Posterior Probability:

$$P(y|\vec{x}) = \left(\prod_{i=1}^{D} P(x_i|y)\right) P(y)$$

Q: Why are we using log probabilities?

A: ... Numerical Underflow ...

Log Posterior Probability:

$$\log P(y|\vec{x}) = \log P(y) + \sum_{i=1}^{D} \log P(x_i|y)$$

Predicted Class: (Highest Posterior Probability)

$$\hat{y} = \arg\max_{y} \log P(y|\vec{x})$$

Logarithmic Rules: $\log(A \times B) = \log A + \log B$

Pseudocode for Gaussian Naïve Bayes

```
# Inputs
#  data     ← (X, y)
#  X_query  ← examples to predict

# ----- fit -----
C ← unique(y)
N, D ← rows(X), cols(X)

FOR each class c ∈ C DO
  Xc     ← rows of X where y = c
  nc     ← rows(Xc)
  π[c]   ← nc / N                          # prior
  μ[c]   ← mean(Xc, axis=0)                # D means
  σ2[c]  ← var(Xc,  axis=0)                # D variances
END FOR

# ----- predict -----
ŷ ← list of length |X_query|

FOR i = 1 TO |X_query| DO
  x* ← X_query[i]
  FOR each c ∈ C DO
    # log-posterior up to a constant
    diff  ← x* - μ[c]
    logp[c] ← log(π[c]) - 0.5 * sum( log(2π·σ2[c]) + (diff◦diff) / σ2[c] )
  END FOR
  ŷ[i] ← argmax_index(logp)                # class with highest score
END FOR

RETURN ŷ
```

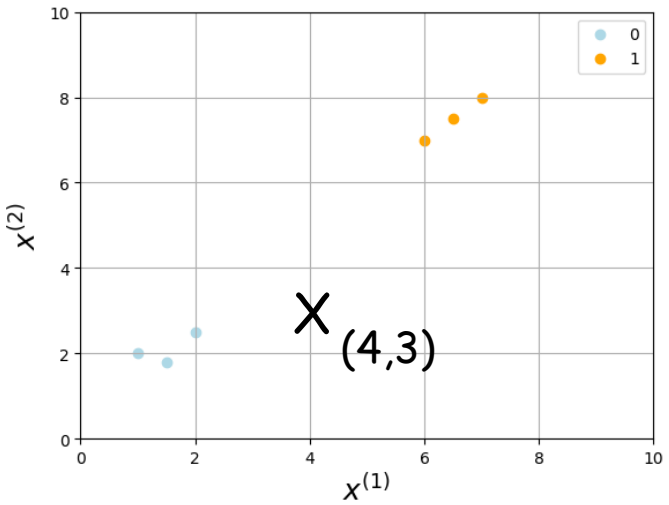
Gaussian Naïve Bayes from Scratch

```
import numpy as np
from sklearn.base import BaseEstimator, ClassifierMixin

class MyGaussianNB(BaseEstimator, ClassifierMixin):
    def __init__(self, var_smoothing=1e-9):
        self.var_smoothing = var_smoothing

    def fit(self, X, y):
        X = np.asarray(X, float); y = np.asarray(y)
        self.classes_, y_idx = np.unique(y, return_inverse=True)
        C = self.classes_.size
        self.class_prior_ = np.bincount(y_idx) / y.size
        self.mean_ = np.vstack([X[y_idx == c].mean(axis=0) for c in range(C)])
        self.var_  = np.vstack([X[y_idx == c].var(axis=0)  for c in range(C)])
        return self

    def predict_proba(self, X):
        X = np.asarray(X, float)
        logp = []
        for prior, mu, var in zip(self.class_prior_, self.mean_, self.var_ + self.var_smoothing):
            ll = -0.5 * (np.log(2*np.pi*var).sum() + ((X - mu)**2 / var).sum(axis=1))
            logp.append(np.log(prior) + ll)
        logp = np.column_stack(logp)
        m = logp.max(axis=1, keepdims=True)
        p = np.exp(logp - m)
        return p / p.sum(axis=1, keepdims=True)

    def predict(self, X):
        return self.classes_[self.predict_proba(X).argmax(axis=1)]
```

Usage Example

```
import numpy as np

# ----- training data -----
X_train = np.array([
    [1.0, 2.0], [1.5, 1.8], [2.0, 2.5],
    [6.0, 7.0], [6.5, 7.5], [7.0, 8.0],
])
y_train = np.array([0, 0, 0, 1, 1, 1])   # classes: 0, 1

# ----- fit Gaussian NB -----
gnb = MyGaussianNB(var_smoothing=1e-9)    # default is fine
gnb.fit(X_train, y_train)

# ----- classify a new point -----
X_test = np.array([[4.0, 3.0]])           # 2D shape (1, 2)
proba = gnb.predict_proba(X_test)[0]      # probabilities in gnb.classes_ order
pred  = gnb.predict(X_test)[0]

# ----- outputs -----
print("Classes order:", gnb.classes_)                # e.g., [0 1]
print(f"New point: {X_test.ravel().tolist()}")
print(f"Probabilities {proba.tolist()}")             # [P(class=0), P(class=1)]
print(f"Predicted class: {int(pred)}")
```

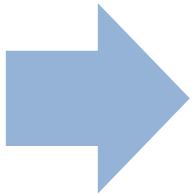
```
Classes order: [0 1]
New point: [4.0, 3.0]
Probabilities [class 0, class 1]: [1.0000000000000000, 3.1923e-25]
Predicted class: 0
```



Step 1: Identifying Unique Classes & Prior Probabilities

```
def fit(self, X, y):
    X = np.asarray(X, float); y = np.asarray(y)

    # Unique Classes
    self.classes_, y_idx = np.unique(y, return_inverse=True)
    C = self.classes_.size

    # Prior Probabilities
    self.class_prior_ = np.bincount(y_idx) / y.size

    self.mean_ = np.vstack([X[y_idx == c].mean(axis=0) for c in range(C)])
    self.var_  = np.vstack([X[y_idx == c].var(axis=0)  for c in range(C)])

    return self
```

$$y = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$



Unique Classes:

$$\begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Prior Probabilities:

$$P\,(y = 0) = \frac{3}{3+3} = 0.5$$

$$P\,(y = 1) = \frac{3}{3+3} = 0.5$$

Step 2: Fitting Conditional Probabilities

```
def fit(self, X, y):
    X = np.asarray(X, float); y = np.asarray(y)
    self.classes_, y_idx = np.unique(y, return_inverse=True)
    C = self.classes_.size
    self.class_prior_ = np.bincount(y_idx) / y.size

    # Mean
    self.mean_ = np.vstack([X[y_idx == c].mean(axis=0) for c in range(C)])

    # Variance → Standard Deviation
    self.var_  = np.vstack([X[y_idx == c].var(axis=0)  for c in range(C)])

    return self
```

$$\vec{x} = \begin{bmatrix} 1 & 2 \\ 1.5 & 1.8 \\ 2 & 2.5 \\ 6 & 7 \\ 6.5 & 7.5 \\ 7 & 8 \end{bmatrix} \quad y = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

Means:

$$\vec{\mu}^{(0)} = [1.5 \quad 2.1]$$

$$\vec{\mu}^{(1)} = [6.5 \quad 7.5]$$

Standard Deviation: $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$

$$\vec{\sigma}^{(0)} = [0.408 \quad 0.294]$$

$$\vec{\sigma}^{(1)} = [0.408 \quad 4.08]$$

Step 3: Calculating Posterior Probabilities

```
def predict_proba(self, X):
    X = np.asarray(X, float)
    logp = []
    for prior, mu, var in zip(self.class_prior_, self.mean_, self.var_ + self.var_smoothing):
        # Log(Posterior Probability)
        ll = -0.5 * (np.log(2*np.pi*var).sum() + ((X - mu)**2 / var).sum(axis=1))

        logp.append(np.log(prior) + ll)

    logp = np.column_stack(logp)
    m = logp.max(axis=1, keepdims=True)
    p = np.exp(logp - m)

    # Normalised Probabilities
    return p / p.sum(axis=1, keepdims=True)
```

$$P(y=0|\vec{x}=(4,3)) = P(x_1=4|y=0)P(x_2=3|y=0)P(y=0)$$
$$= 6.875\times10^{-9}\times1.252\times10^{-2}\times0.5$$
$$= 4.451\times10^{-11}$$

$$P(y=1|\vec{x}=(4,3)) = P(x_1=4|y=1)P(x_2=3|y=1)P(y=1)$$
$$= 6.875\times10^{-9}\times3.756\times10^{-27}\times0.5$$
$$= 1.421\times10^{-35}$$

Normalised Probabilities

$$\frac{4.451\times10^{-11}}{4.451\times10^{-11}+1.421\times10^{-35}} = 0.999\ldots$$

$$\frac{1.421\times10^{-35}}{4.451\times10^{-11}+1.421\times10^{-35}} = 3.19\times10^{-25}$$

Step 4: Class Prediction

```
def predict(self, X):
    return self.classes_[self.predict_proba(X).argmax(axis=1)]
```

$$\arg\max_{y} \{0.99 \ldots, 3.19 \times 10^{-25}\} = 0$$

Scikit-learn: Gaussian Naïve Bayes

```
import numpy as np
from sklearn.naive_bayes import GaussianNB

# training data
X_train = np.array([[1.0, 2.0], [1.5, 1.8], [2.0, 2.5],
                    [6.0, 7.0], [6.5, 7.5], [7.0, 8.0]])
y_train = np.array([0, 0, 0, 1, 1, 1])  # classes: 0, 1

# test point
X_test = np.array([[4.0, 3.0]])

# train & predict
gnb = GaussianNB()              # default var_smoothing is fine
gnb.fit(X_train, y_train)

proba = gnb.predict_proba(X_test)[0]  # probabilities in gnb.classes_ order
pred  = gnb.predict(X_test)[0]

print("Classes order:", gnb.classes_)                # → [0 1]
print("Test Point:", X_test[0].tolist())
print("Probabilities [class 0, class 1]:", proba.tolist())
print("Predicted Class:", int(pred))
```


```
Classes order: [0 1]
New point: [4.0, 3.0]
Probabilities [class 0, class 1]: [1.0000000000000000, 3.1923e-25]
Predicted class: 0
```

Smoothing Variable: Numerical Stability

Conditional Probability:

$$P(x_i|y=c) = N(\mu_i, \sigma_i)$$

$$= \frac{1}{\sqrt{2\pi \times \sigma_i^2}} e^{\left(-\frac{(x-\mu_i)^2}{2\times\sigma_i^2}\right)}$$

If $\sigma_i \to 0$, then $-\frac{(x-\mu_i)^2}{2\times\sigma_i^2}$ will be divided by (near) zero.

Hence, causing a numerical instability.

Smoothing Variable ε:

$$P(x_i|y=c) = \frac{1}{\sqrt{2\pi \times (\sigma_i^2 + \varepsilon)}} e^{\left(-\frac{(x-\mu_i)^2}{2\times(\sigma_i^2+\varepsilon)}\right)}$$

$\varepsilon > 0$ is a small constant. It is to prevent division by (near) zero.

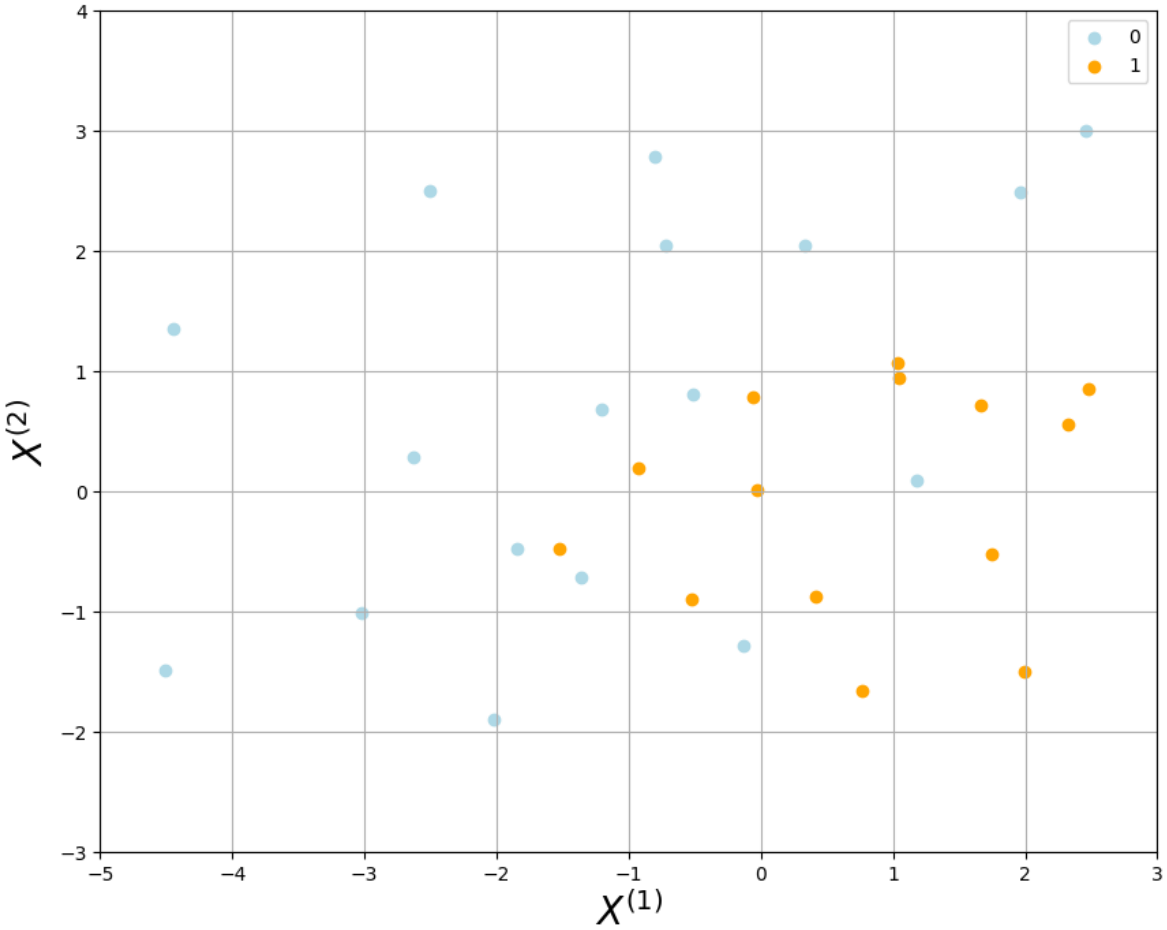
scikit-learn's GaussianNB uses var_smoothing (default 10^{-9}) to set ε as a fraction of the overall feature variance:

$$\varepsilon = \text{var\_smooting} \times \max_j \text{Var}(x_j)$$

i.e. we add 10^{-9} times the **largest** feature variance.

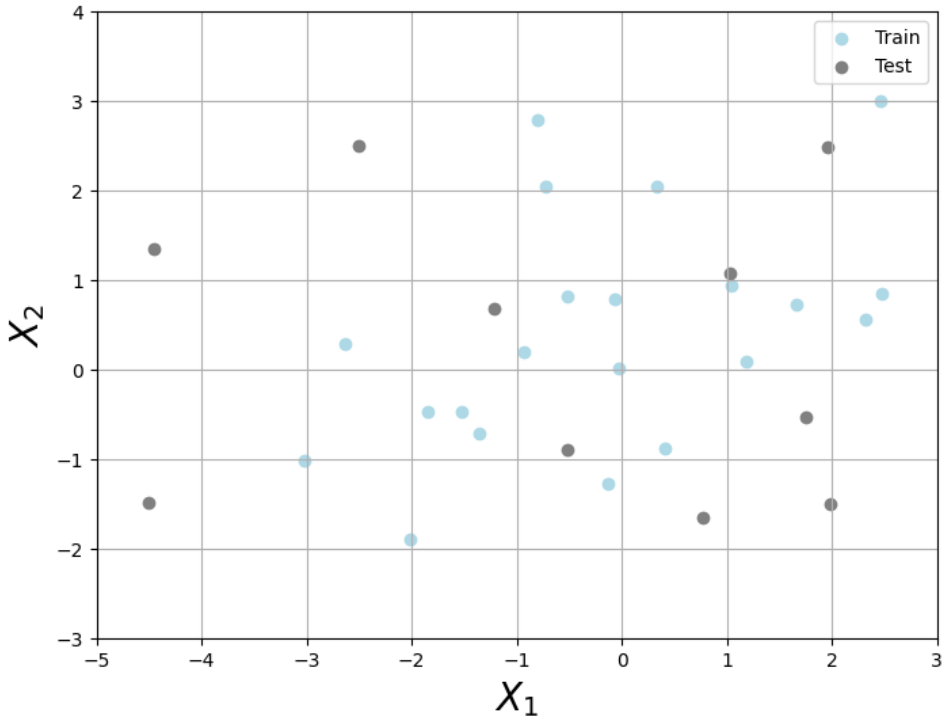
In Gaussian NB, ε ("var_smoothing") is primarily for numerical stability, not a core modeling knob like α in Multinomial NB.

Example Dataset



Train, Validation and Test Datasets

```
from sklearn.model_selection import train_test_split

# First, split the data into train (70%) and test (30%)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42, stratify=y)
```



- Train:Test is typicaly either 0.8:0.2 or 0.7:0.3.
- **Test** dataset is a proxy of unseen data, and it will only be used in the final evaluation.
- **Train** dataset is further divided into k folds (e.g., 5 or 10).
- What hyper-parameters are we optimizing? **None for Gaussian Naïve Bayes.** Hence, k-fold cross-validation is somewhat irrelevant in the Gaussian Naïve Bayes context.

`var_smoothing` *can* be treated as a (usually low-impact) hyperparameter.

- When to tune: small datasets, near-constant features, ill-scaled features.
- When not to bother: features standardised and class variances are well-behaved—default usually fine.

Decision Boundary

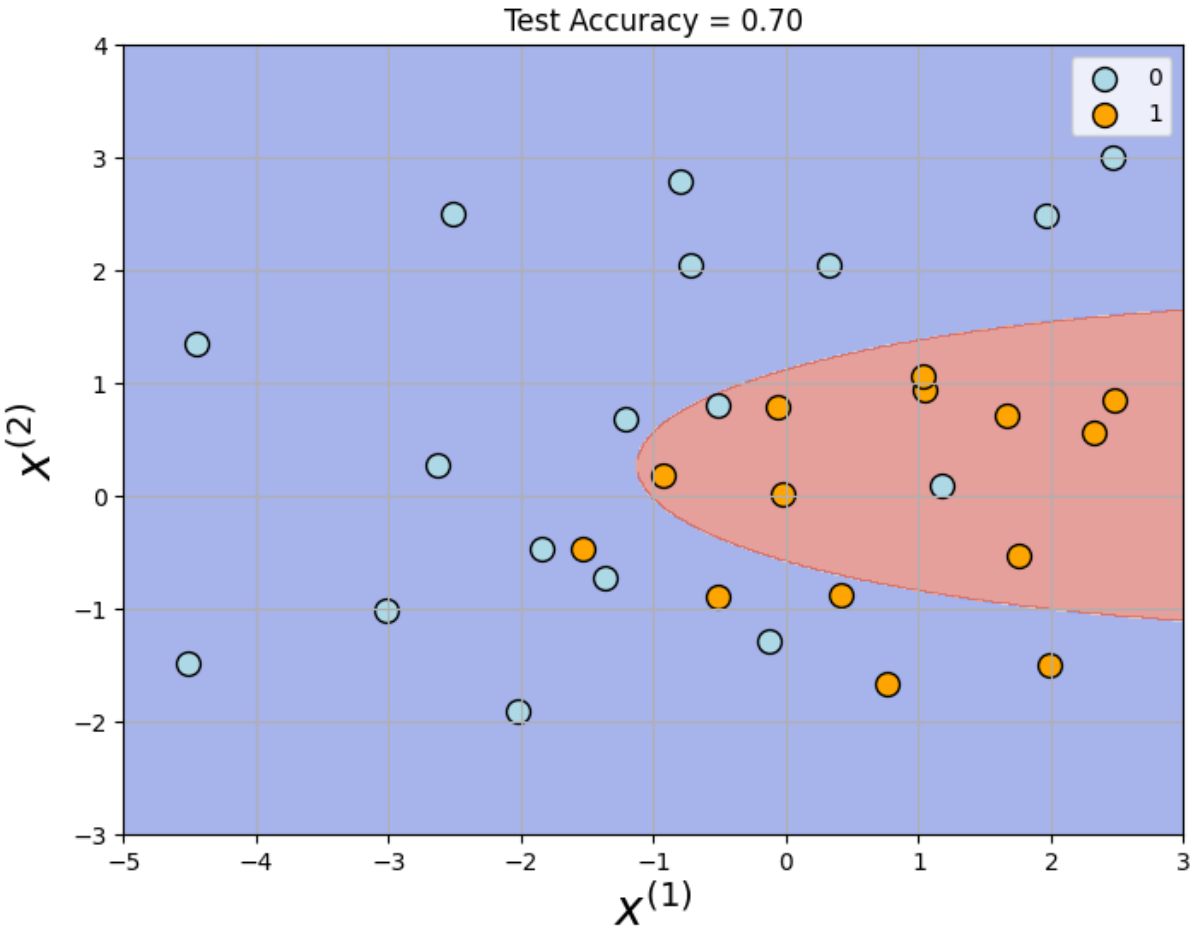
Q: Can accuracy performance of Naïve Bayes be better?

A: No.

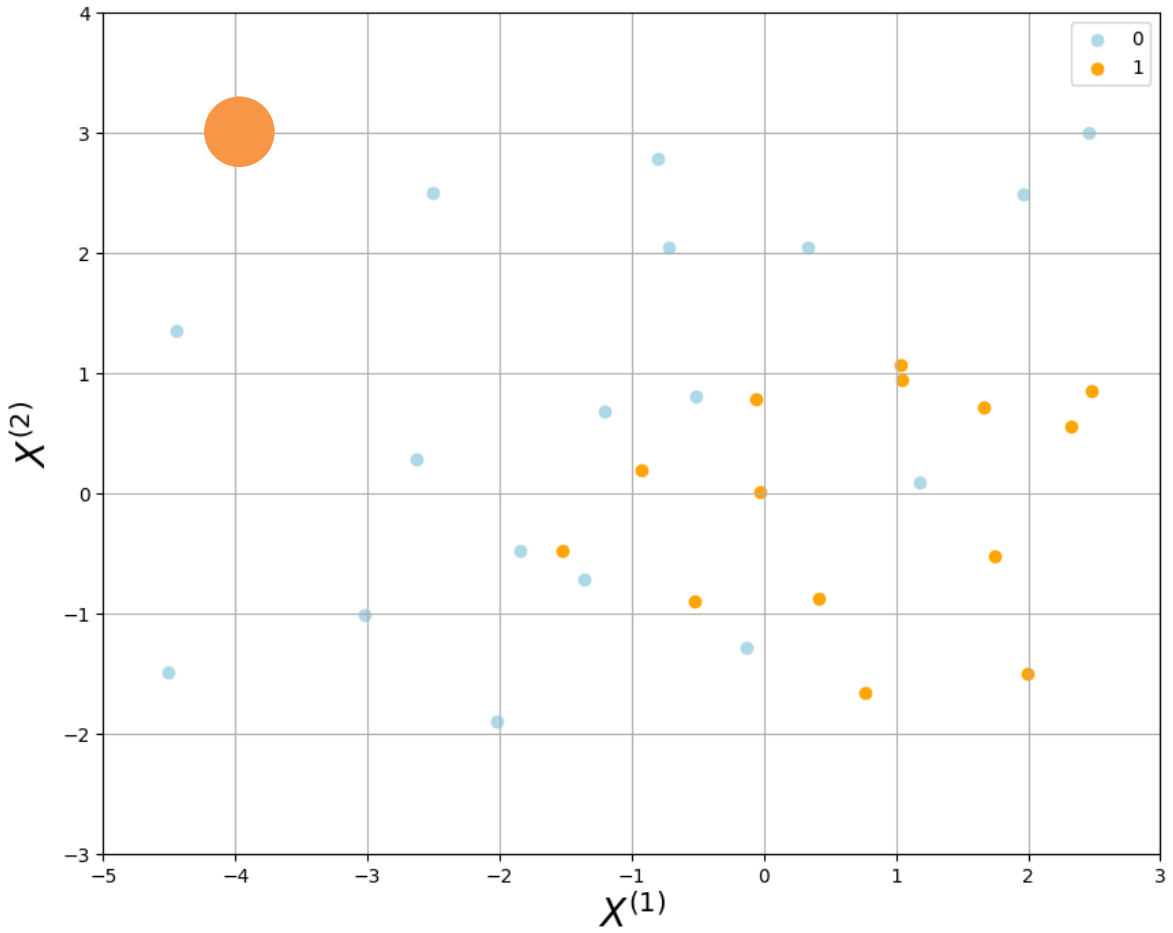
Q: Can we worse than this?

A: (Also) No.

The model simply captures underlying statistics in the data as is. That's all it does. It makes no effort to predict correctly.

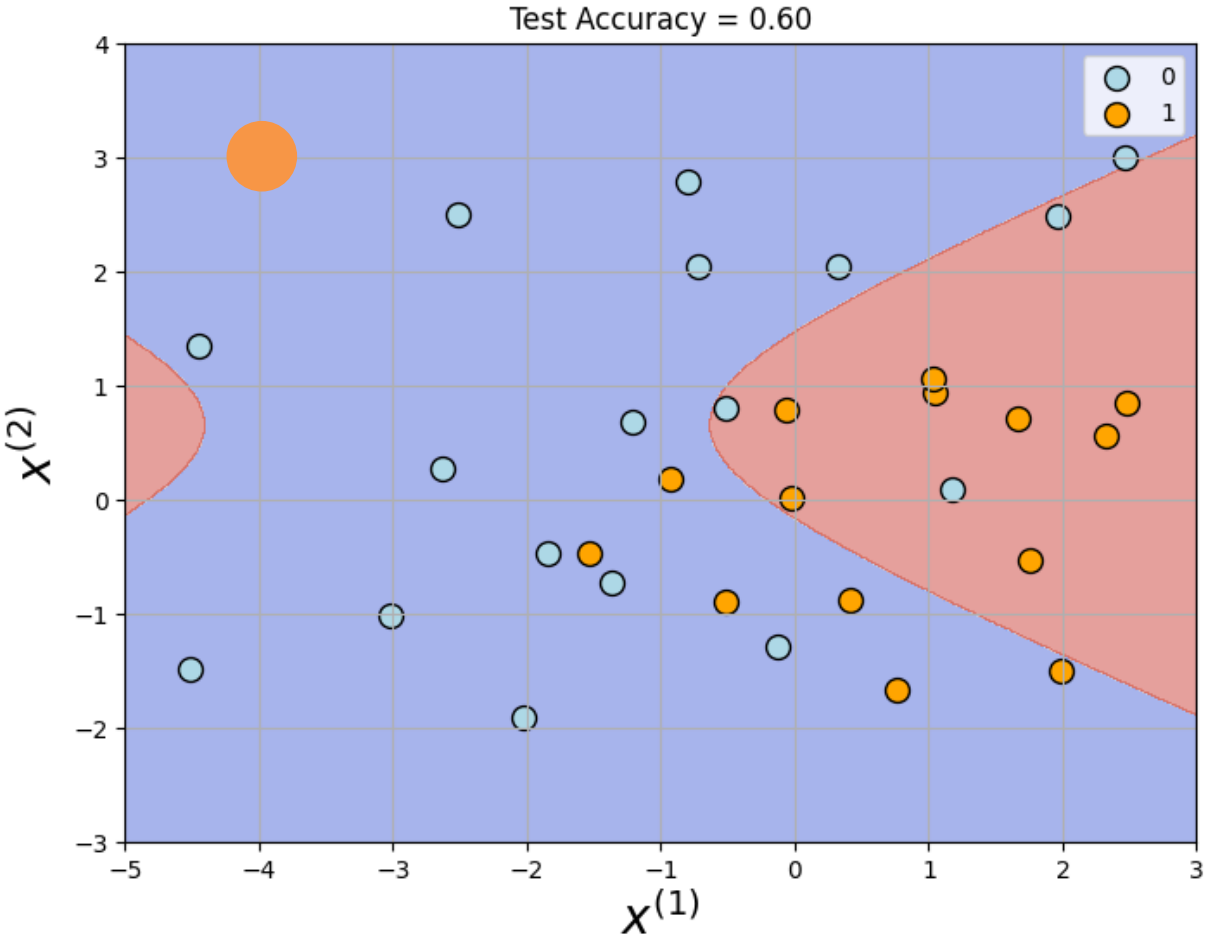


Outliers



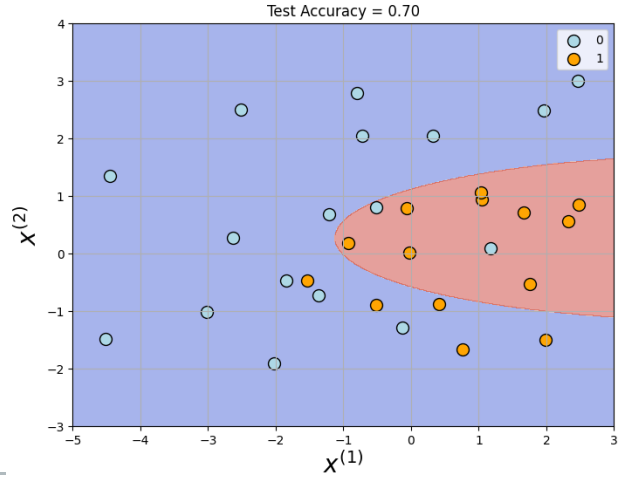
Outliers can distort underlying statistics by significantly altering mean and variance of the conditional probabilities. The model tries to fit these extreme points rather than the overall data trend.

Outliers (cont.)



Q: What are the underlying causes making our Naïve Bayes performs poorly?

A: We assume our data are normally distributed. Mean and variance are not statistically robust.



Multinomial Naïve Bayes

Spam

Free cash now!
Limited offer!
Cash prize waiting!

Ham

We meet tomorrow, can we?
Have you seen my book?
I will bring cash later.

$P(\text{Spam}|\text{"Limited Cash Offer Now!"})$

$P(\text{Ham}|\text{"Limited Cash Offer Now!"})$

Posterior Probability

Q: How do we calculate these?

A: Multinomial Naïve Bayes

Document Representation: Bag of Words

Spam

Free cash now!
Limited offer!
Cash prize waiting!

Ham

We meet tomorrow, can we?
Have you seen my book?
I will bring cash later.

>10k is common for a vocabulary size.

Unique Words in Corpus

Vocabulary = {book, bring, can, cash, free, have, i, later, limited, meet, my, now, offer, prize, seen, tommorow, waitng, we, will, you }

	book	bring	can	cash	free	have	i	later	limited	meet	my	now	offer	prize	seen	tomorrow	waiting	we	will	you
Free cash now!	0	0	0	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
Limited offer!	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0
Cash prize waiting!	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0
We meet tomorrow, can we?	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	1	0	2	0	0
Have you seen my book?	1	0	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	1
I will bring cash later.	0	1	0	1	0		1	1	0	0	0	0	0	0	0	0	0	0	1	0

Features ($\vec{x}$): M-dimensional vector, where M equals the number of words in the dictionary.

Multinomial Naïve Bayes

$$P(\text{Spam}|\text{offer}, \text{free}, \ldots, \text{cash}) \propto P(\text{offer}|\text{Spam})P(\text{free}|\text{Spam}) \ldots P(\text{cash}|\text{Spam})P(\text{Spam})$$

New Email

Conditions

Q: How do we calculate these?

A: Multinomial Naïve Bayes

Prior Probability:

$$P(\text{Spam}) = \frac{\text{Number of Spam Emails}}{\text{Total Number of Emails}}$$

Conditional Probability:

$$P(\text{offer}|\text{Spam}) = \frac{\text{Count}(\text{offer}|\text{Spam}) + \alpha}{\text{Total Words in Spam Emails} + \alpha|V|}$$

Count(offer|Spam) is 'How many times does the token "offer" appear in all training e-mails that are labelled Spam?'.

It is <u>not</u> the number of spam messages that contain "offer".

Why Multinomial NB uses Token counts instead of document frequency?

Perspective	Token-Count Likelihood (Multinomial NB)	Document-Frequency Likelihood (Bernoulli NB)
What is modelled?	Every *occurrence* of a word is treated as an independent draw from the class-specific distribution $P(w\|c)$.	Only the *presence/absence* of the word in a document is modelled, via $P(\text{appears} = 1\|c)$. Repeated hits add **no** extra evidence.
Sufficient Statistic	Aggregate **token counts** $N_{\text{ic}} = \sum_{d \epsilon c} n_{\text{id}}$.	**Document counts** $df_{\text{ic}} = \{d \epsilon c : n_{\text{id}} > 0\}$.
Signal Retained	Strength of evidence grows with repetition ("offer offer offer" is louder than a single "offer").	All repetitions collapse to one bit.
Best For	Longer texts, bag-of-words spam/news/topic tasks, streaming counts.	Very short texts (tweets, subject lines), IR binary vector space, situations where length bias must be zero.

Example (Spam vs Ham):

Email	Text	"offer" Count
A (Ham)	"... attached is our **offer** letter for your internship ..."	1
B (Spam)	"**OFFER OFFER OFFER**! Limited-time offer now!"	3

Conditional Probability

Spam	Ham
Free cash now! Limited offer! Cash prize waiting!	We meet tomorrow, can we? Have you seen my book? I will bring cash later.

Vocabulary = {book, bring, can, cash, free, have, i, later, limited, meet, my, now, offer, prize, seen, tommorow, waitng, we, will, you }

$$P(\text{cash}|\text{Spam}) = \frac{\text{Count}(\text{cash}|\text{Spam}) + \alpha}{\text{Total Words in Spam Emails} + \alpha|V|}$$

$$= \frac{2+1}{8+1\cdot 20} = \frac{3}{28}$$

$$P(\text{cash}|\text{Ham}) = \frac{\text{Count}(\text{cash}|\text{Ham}) + \alpha}{\text{Total Words in Ham Emails} + \alpha|V|}$$

$$= \frac{1+1}{15+1\cdot 20} = \frac{2}{35}$$

$|V|$: Vocaborary Size

$\propto$: Laplace Smoothing (usually=1)

$\propto$ is for handling zero probabilities for words not seen the training data.

Limited Cash Offer Now! Free Cash!

$$P(y=0|\text{"Limited Cash Offer Now! Free Cash!"}) = P(\text{limited}|y=0)\times P(\text{cash}|y=0)^2\times P(\text{offer}|y=0)$$

$$\times P(\text{now}|y=0)\times P(\text{free}|y=0)\times P(y=0)$$

$$= \frac{1}{35}\times\left(\frac{2}{35}\right)^2\times\frac{1}{35}\times\frac{1}{35}\times\frac{1}{35}\times\frac{1}{2}$$

$$= \frac{2}{35^6}$$

$$= 1.088\times10^{-9}$$

Prior Probability:

$$P(y=0) = \frac{\text{Number of Ham Emails}}{\text{Number of Ham Emails} + \text{Number of Spam Emails}}$$

$$= \frac{3}{3+3}$$

$$= 0.5$$

Spam (1)

Free cash now!
Limited offer!
Cash prize waiting!

Ham (0)

We meet tomorrow, can we?
Have you seen my book?
I will bring cash later.

Posterior Probability (cont.)

Limited Cash Offer Now! Free Cash!

$$P(y=1|\text{"Limited Cash Offer Now! Free Cash!"}) = P(\text{limited}|y=1)\times P(\text{cash}|y=1)^2\times P(\text{offer}|y=1)$$
$$\times P(\text{now}|y=1)\times P(\text{free}|y=1)\times P(y=1)$$
$$= \frac{2}{28}\times\left(\frac{3}{28}\right)^2\times\frac{2}{28}\times\frac{2}{28}\times\frac{2}{28}\times\frac{1}{2}$$
$$= \frac{2^3\times 3^2}{28^6}$$
$$= 1.494\times 10^{-7}$$

Prior Probability:

$$P(y=1) = \frac{\text{Number of Spam Emails}}{\text{Number of Ham Emails} + \text{Number of Spam Emails}}$$
$$= \frac{3}{3+3}$$
$$= 0.5$$

Spam (1)

Free cash now!
Limited offer!
Cash prize waiting!

Ham (0)

We meet tomorrow, can we?
Have you seen my book?
I will bring cash later.

Prediction: Highest Posterior Probability

$$P(y = 0|\text{"Limited Cash offer Now! Free Cash!"}) = 1.088\times10^{-9}$$

$$P(y = 1|\text{"Limited Cash offer Now! Free Cash"}) = 1.494\times10^{-7}$$

Since $1.494\times10^{-7} > 1.088\times10^{-9}$, "Limited Cash Offer Now! Free Cash!" is predicted to belong to class 1, i.e. Spam.

Limited Cash Offer Now! Free Cash!

Pseudocode for Multinomial Naïve Bayes

```
# Inputs
#  data     ← (X, y) with nonnegative integer features (counts)
#  α        ← additive smoothing (e.g., α = 1)
#  X_query  ← examples to predict

# ----- fit -----
C, N, D ← unique(y), rows(X), cols(X)

FOR each class c ∈ C DO
  Xc        ← rows of X where y = c
  nc        ← rows(Xc)
  π[c]      ← nc / N                                  # prior P(y=c)
  count[c]  ← sum(Xc, axis=0)                         # D-length count vector
  total[c]  ← sum(count[c])                           # total tokens in class c
  θ[c, j]   ← (count[c][j] + α) / (total[c] + α·D)    # P(feature j | y=c)
END FOR

# ----- predict -----
ŷ ← list of length |X_query|

FOR i = 1 TO |X_query| DO
  x* ← X_query[i]
  FOR each c ∈ C DO
    # log-posterior up to a constant (use logs for stability)
    logp[c] ← log(π[c]) + Σ_{j=1..D} x*[j] · log(θ[c, j])
  END FOR
  ŷ[i] ← argmax_index(logp)                           # class with highest score
END FOR

RETURN ŷ
```

Python Code Snippet

```
import numpy as np
from sklearn.base import BaseEstimator, ClassifierMixin

class MyMultinomialNB(BaseEstimator, ClassifierMixin):
    def __init__(self, alpha: float = 1.0):
        self.alpha = alpha

    def fit(self, X, y):
        X = np.asarray(X, float); y = np.asarray(y)
        self.classes_, y_idx = np.unique(y, return_inverse=True)
        C, D = self.classes_.size, X.shape[1]

        # priors
        class_count = np.bincount(y_idx, minlength=C).astype(float)
        self.class_log_prior_ = np.log(class_count / class_count.sum())

        # feature counts per class
        feature_count = np.zeros((C, D), dtype=float)
        for c in range(C):
            feature_count[c] = X[y_idx == c].sum(axis=0)

        # additive smoothing
        alpha = float(self.alpha)
        smoothed = feature_count + alpha
        class_feature_sum = smoothed.sum(axis=1, keepdims=True)  # total tokens per class
        self.feature_log_prob_ = np.log(smoothed / class_feature_sum)

        return self

    def _joint_log_likelihood(self, X):
        X = np.asarray(X, float)
        # log P(y=c) + sum_j x_j * log P(feature_j | c)
        return X @ self.feature_log_prob_.T + self.class_log_prior_
```

```
…cont…

def predict_proba(self, X):
    logp = self._joint_log_likelihood(X)
    m = logp.max(axis=1, keepdims=True)
    p = np.exp(logp - m)
    return p / p.sum(axis=1, keepdims=True)

def predict(self, X):
    logp = self._joint_log_likelihood(X)
    idx = np.argmax(logp, axis=1)
    return self.classes_[idx]
```

Usage Example

```
import numpy as np
# from your_module import MyMultinomialNB   # if the class isn't in the same file

# Vocabulary (implicitly):
# Index:  0      1     2    3      4       5    6   7   8   9    10   11   12  13   14
# Word:  [limited, cash, now, offer, prize, waiting, can, we, meet, have, you, seen, i, will, bring]

X_train = np.array([
    [1,1,1,0,0,0,0,0,0,0,0,0,0,0,0],  # "limited cash now"     → Spam
    [0,0,0,1,1,1,0,0,0,0,0,0,0,0,0],  # "offer prize waiting"  → Spam
    [0,2,0,0,0,0,0,0,0,0,0,0,0,0,0],  # "cash cash"            → Spam
    [0,0,0,0,0,0,1,1,1,0,0,0,0,0,0],  # "can we meet"          → Ham
    [0,0,0,0,0,0,0,0,0,1,1,1,0,0,0],  # "have you seen"        → Ham
    [0,1,0,0,0,0,0,0,0,0,0,0,1,1,1]   # "i will bring cash"    → Ham
])

y_train = np.array([1, 1, 1, 0, 0, 0])   # 1 = Spam, 0 = Ham

# New sample: "limited cash offer"  → BoW: limited=1, cash=1, offer=1
new_sample = np.array([[1,1,0,1,0,0,0,0,0,0,0,0,0,0,0]])  # 2D shape (1, D)

# Train Multinomial NB (Laplace smoothing alpha=1.0)
model = MyMultinomialNB(alpha=1.0)
model.fit(X_train, y_train)

# Predict
proba = model.predict_proba(new_sample)[0]   # [P(Ham), P(Spam)] since classes_ = [0, 1]
pred  = model.predict(new_sample)[0]

print("Classes order:", model.classes_)      # expect [0 1]
print(f"Probabilities  [Ham, Spam]: {proba}")
print(f"Predicted Class: {pred} ({'Spam' if pred == 1 else 'Ham'})")
```

```
Classes order: [0 1] (0 = Ham, 1 = Spam)
Probabilities [Ham, Spam]: [0.0887, 0.9113]
Predicted Class: 1 (Spam)
```

Python Code Snippet

$$y = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \Rightarrow \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Python Code Snippet

$$y = [1 \quad 1 \quad 1 \quad 0 \quad 0 \quad 0]^T$$

Priors: $P(y = 1) = \frac{3}{3+3} = 0.5$

$$P(y = 0) = \frac{3}{3+3} = 0.5$$

$$\text{Count}(1) = [1 \quad 3 \quad 1 \quad 1 \quad 1 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0]^T \rightarrow \text{Count}_{\text{Total}}(1) = 8$$

$$\text{Count}(0) = [0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1]^T \rightarrow \text{Count}_{\text{Total}}(0) = 10$$

Python Code Snippet (cont.)

$$P(\cdot|1) = \frac{\text{Count}(\cdot|1) + \alpha}{\text{Total Words in Class } 1 + \alpha|V|}$$

$$P(\cdot|0) = \frac{\text{Count}(\cdot|0) + \alpha}{\text{Total Words in Class } 0 + \alpha|V|}$$

$$P(\ldots|1) = \left[\frac{2}{23} \quad \frac{4}{23} \quad \frac{2}{23} \quad \frac{2}{23} \quad \frac{2}{23} \quad \frac{2}{23} \quad \frac{1}{23} \quad \frac{1}{23} \quad \frac{1}{23} \quad \frac{1}{23} \quad \frac{1}{23} \quad \frac{1}{23} \quad \frac{1}{23} \quad \frac{1}{23} \quad \frac{1}{23} \quad \frac{1}{23}\right]^T$$

$$P(\ldots|0) = \left[\frac{1}{25} \quad \frac{2}{25} \quad \frac{1}{25} \quad \frac{1}{25} \quad \frac{1}{25} \quad \frac{1}{25} \quad \frac{2}{25} \quad \frac{2}{25} \quad \frac{2}{25} \quad \frac{2}{25} \quad \frac{2}{25} \quad \frac{2}{25} \quad \frac{2}{25} \quad \frac{2}{25} \quad \frac{2}{25} \quad \frac{2}{25}\right]^T$$

$$P(y = 1|\text{"Limited Cash Offer!"}) = \frac{2}{23} \times \frac{4}{23} \times \frac{2}{23} \times \frac{1}{2} = 6.575 \times 10^{-4}$$

$$P(y = 0|\text{"Limited Cash Offer!"}) = \frac{1}{25} \times \frac{2}{25} \times \frac{1}{25} \times \frac{1}{2} = 6.400 \times 10^{-5}$$

Python Code Snippet (cont.)

Unnormalised Joint Scores:

$$P(y = 0|\text{"Limited Cash Offer!"}) = 6.400\times10^{-5}$$

$$P(y = 1|\text{"Limited Cash Offer!"}) = 6.575\times10^{-4}$$

Normalised Posteriors:

$$P(y = 0|\text{"Limited Cash Offer!"}) = 0.0887$$

$$P(y = 1|\text{"Limited Cash Offer!"}) = 0.9113$$

$0.9113 > 0.0887$, hence "Limited Cash Offer!" is 1, i.e. Spam.

Scikit-learn: Multinomial Naïve Bayes

```
import numpy as np
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.naive_bayes import MultinomialNB

# ----- data -----
texts = [
    "limited cash now", "offer prize waiting", "cash cash",
    "can we meet tomorrow", "have you seen my book", "I will bring cash later" # Ham
]
y = np.array([1, 1, 1, 0, 0, 0])  # 1 = Spam, 0 = Ham

# ----- vectorize -----
vec = CountVectorizer()
X = vec.fit_transform(texts)
vocab = vec.get_feature_names_out()  # index order used by the model

# ----- train on all data -----
clf = MultinomialNB(alpha=1.0)       # Laplace smoothing α=1
clf.fit(X, y)

# ----- score a new sample -----
new_msg = ["limited cash offer"]
X_new = vec.transform(new_msg)
proba_new = clf.predict_proba(X_new)[0]
pred_new  = clf.predict(X_new)[0]

print("Vocabulary (index order):", list(vocab))
print("Predicted classes on training set:", y_pred_train.tolist())
print(f'New: "{new_msg[0]}"  →  probs [ham, spam] = {proba_new.round(4).tolist()},  pred = {int(pred_new)}')
print("Classes order:", clf.classes_)     # should be [0 1]
```

```
Vocabulary (index order): ['book', 'bring', 'can', 'cash', 'have', 'later', 'limited',
                           'meet', 'my', 'now', 'offer', 'prize', 'seen',
                           'tomorrow', 'waiting', 'we', 'will', 'you']
Predicted classes on training set: [1, 1, 1, 0, 0, 0]
New: "limited cash offer"  →  probs [ham, spam] = [0.0687, 0.9313],  pred = 1
Classes order: [0 1]
```

Numerical Underflow

$P(x^{(i)}|y)$ are less than 1. $P(y|\vec{x}) = \prod_{i=1}^{|V|} P(x^{(i)}|y)$, hence $P(y|\vec{x}) \rightarrow 0$.

This could lead to numerical underflow. *Try: 0.1**1000 (i.e. 1.0×10^{-1000}) in Python.*

Q: How can we handle very very small numbers?

A: We take Log. 1000xLog(0.1) = 1,000x(-1) = -1,000.

Log Likelihood Function:

$$\log P(y|\vec{x}) \quad = \quad \log(y) + \sum_{i=1}^{|V|} \log P(x^{(i)}|y)$$

Stanford's IMDB Dataset

50k Movie Reviews

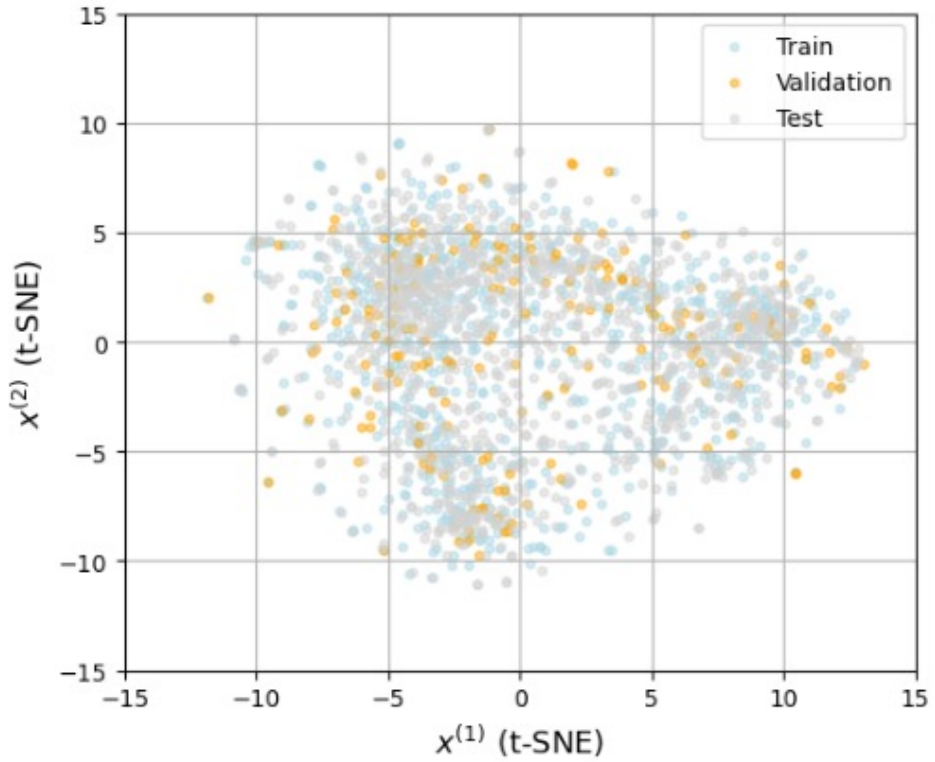
- 25k for Train and 25k for Test
- 2 Classes, Positive and *Negative* Sentiments
- 44,490 Vocabulary Size

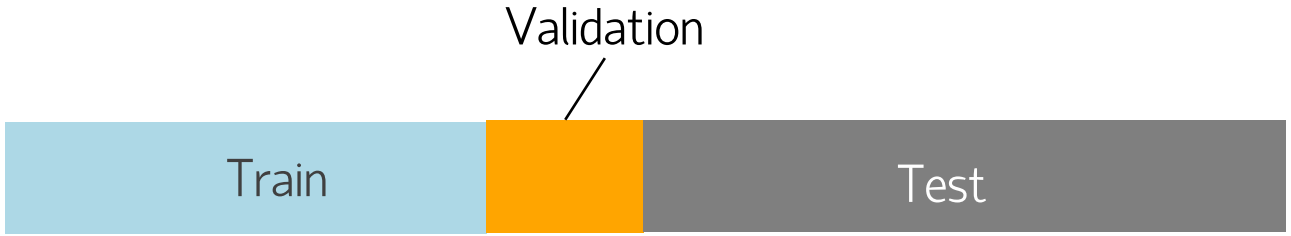

```
If you like adult comedy cartoons, like South Park, then this is nearly a similar
format about the small adventures of three teenage girls at Bromwell High.
Keisha, Natella and Latrina have given exploding sweets and behaved like bitches,
I think Keisha is a good leader. There are also small stories going on with the
teachers of the school. There's the idiotic principal, Mr. Bip, the nervous
Maths teacher and many others. The cast is also fantastic, Lenny Henry's Gina
Yashere, EastEnders Chrissie Watts, Tracy-Ann Oberman, Smack The Pony's Doon
Mackichan, Dead Ringers' Mark Perry and Blunder's Nina Conti. I didn't know this
came from Canada, but it is very good. Very good!
```

https://ai.stanford.edu/~amaas/data/sentiment/

Train, Validation and Test Datasets

```
from sklearn.model_selection import train_test_split

# First, split the data into train (80%) and validation (20%)
X_train, X_val, y_train, y_val= train_test_split(X_train, y_train, test_size=0.2, random_state=42, stratify=y)
```

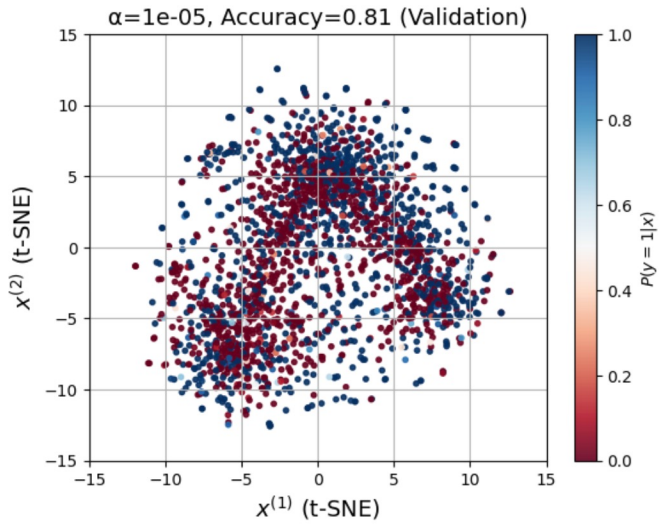


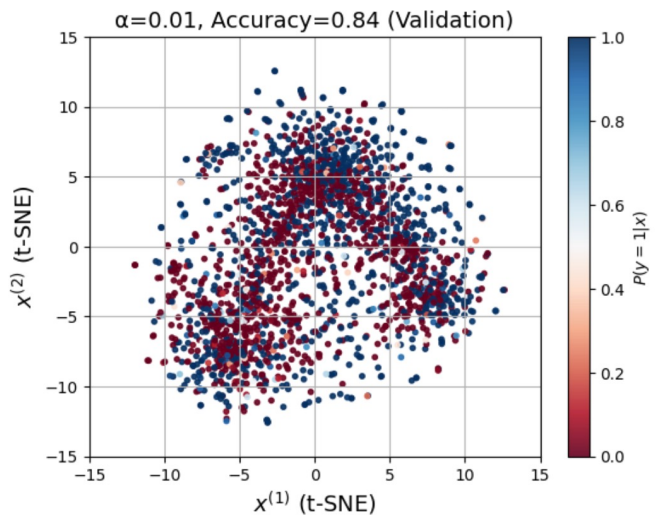


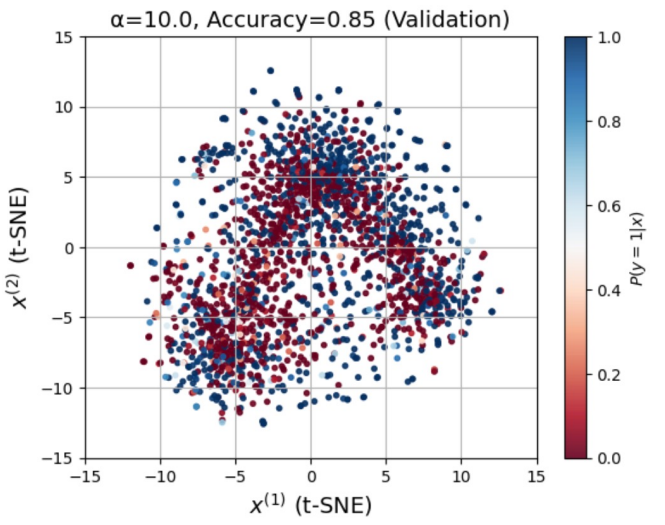
- (Train:Validation):Test is (0.4:0.1):0.5.

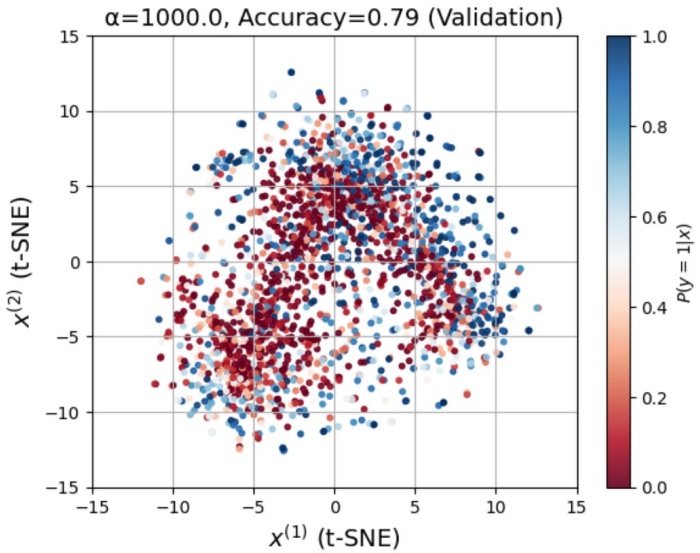
- **Test** dataset is a proxy of unseen data, and it will only be used in the final evaluation.
- We train our ML model on the **Train** dataset.
- **Validation** datest is used to fine-tune or optimise the ML model. Here, we find a smoothing alpha α that will result in the best performance on the Validation dataset.

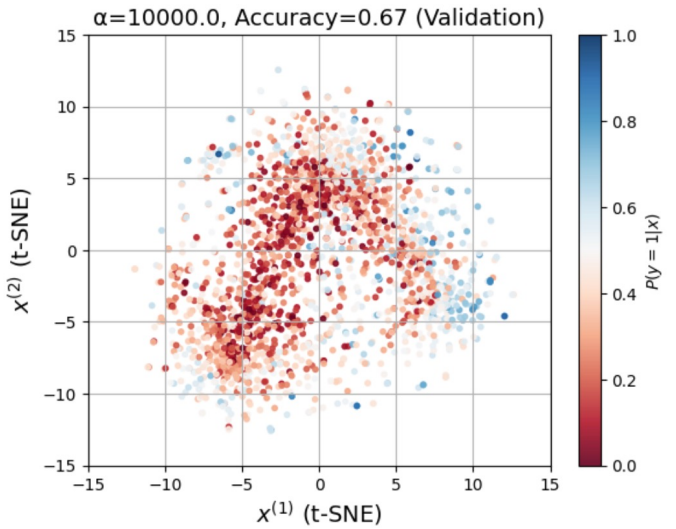
Model Complexity

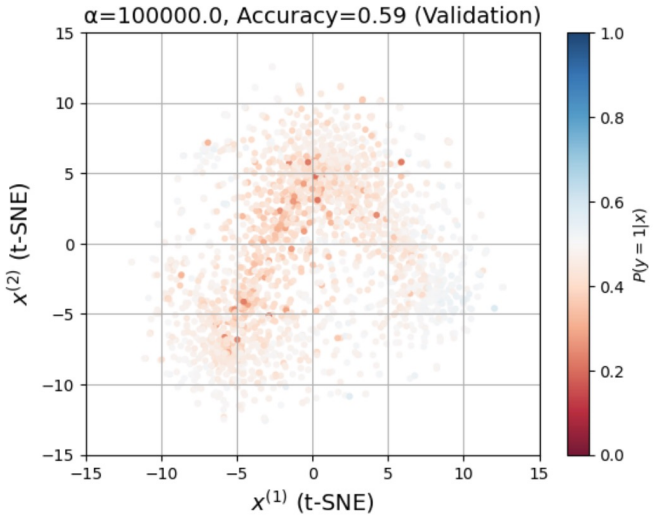




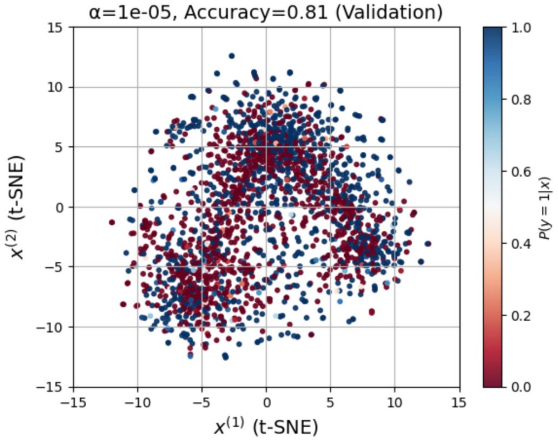




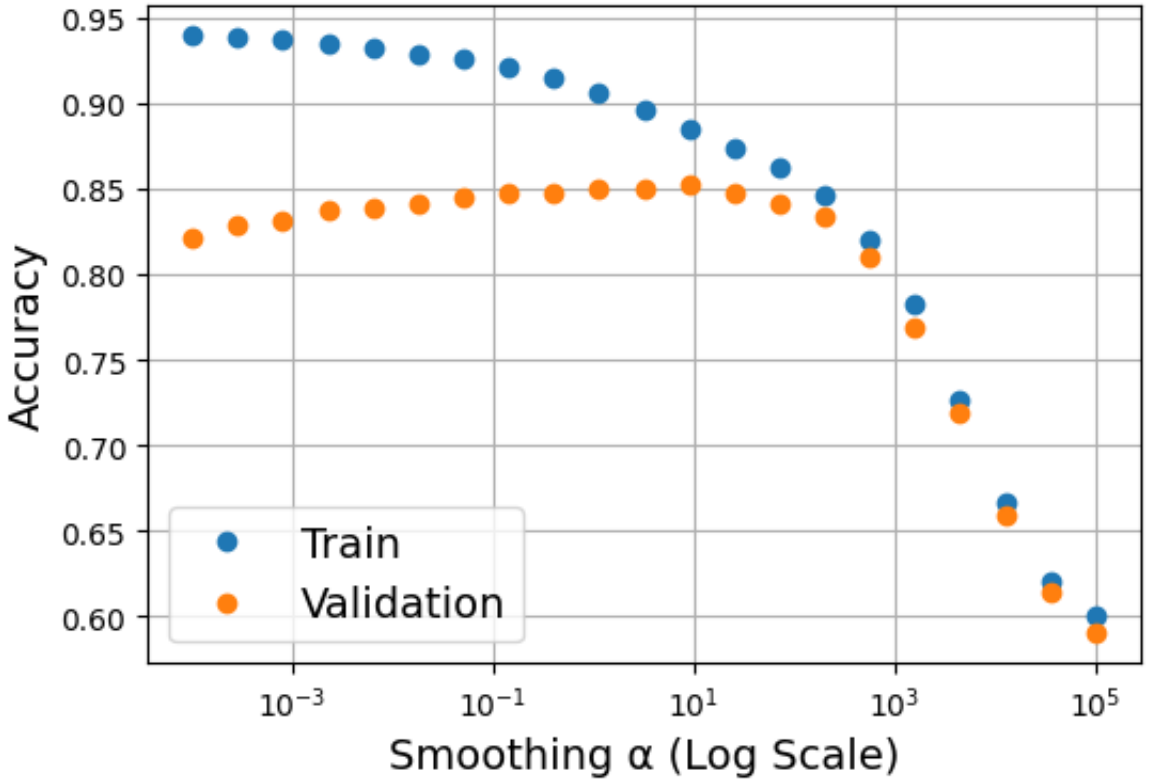


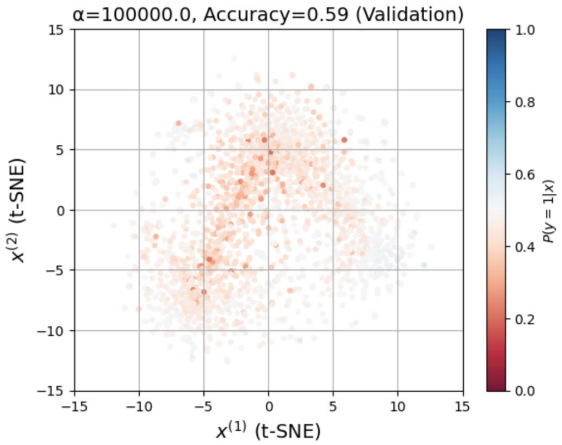


Bias-Variance Trade-Offs



- High Variance, Low Bias
- Complex Probability Density
- High Train Accuracy
- Low Validation Accuracy

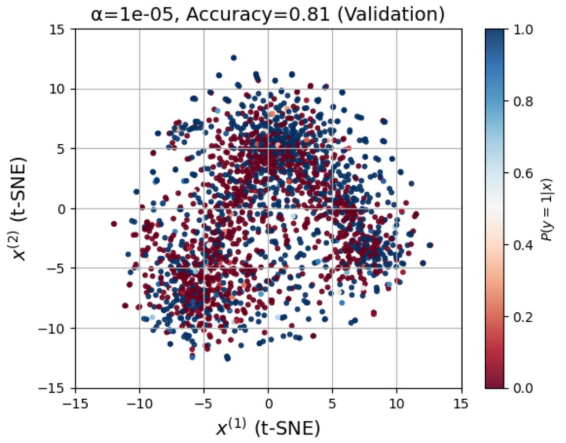




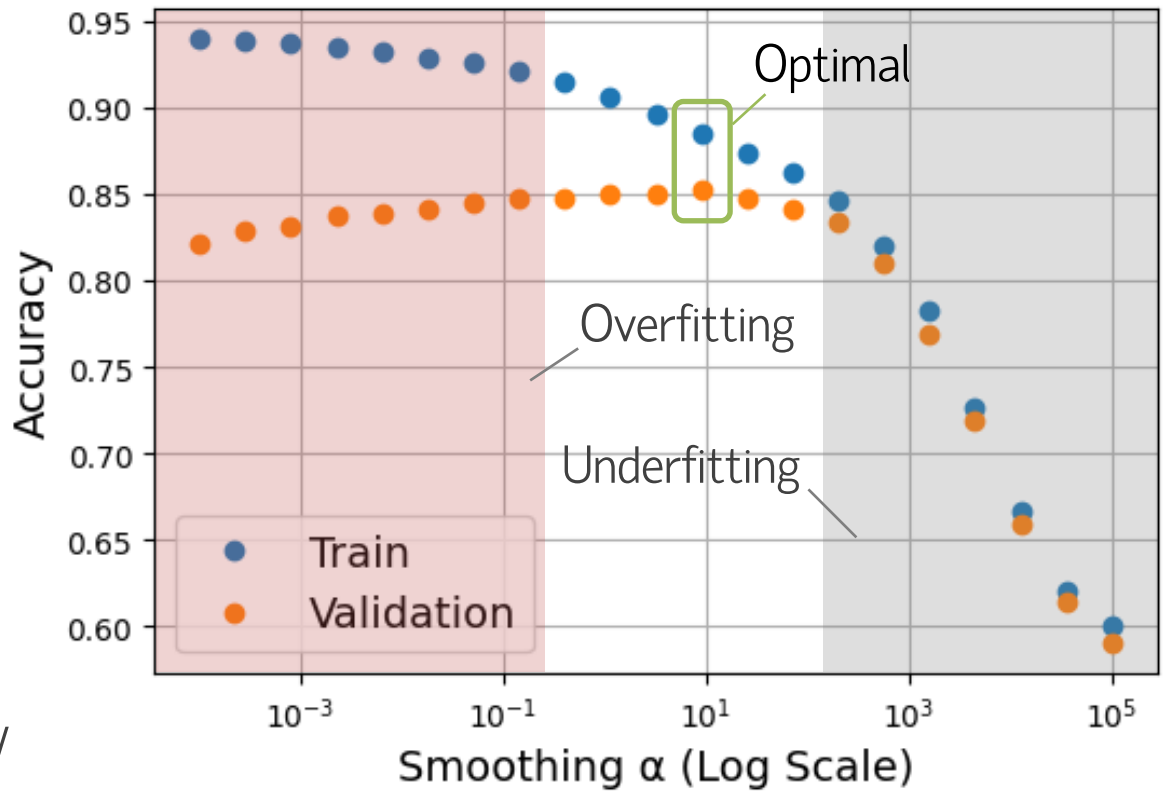
- High Bias, Low Variance
- Simple Probability Density
- Low Train Accuracy
- Low Validation Accuracy

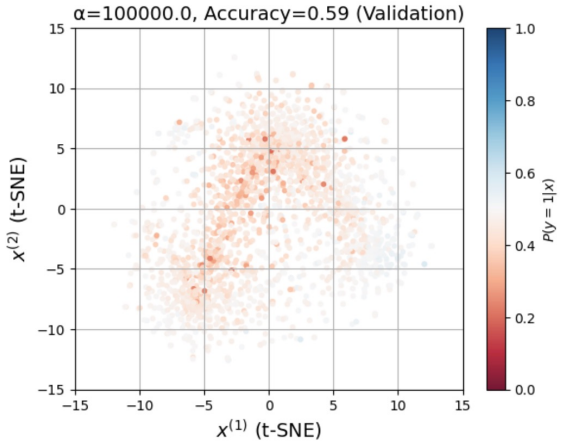
Overfitting vs Underfitting

- **Overfitting** happens when a small smoothing alpha α leads to high training accuracy but poor validation accuracy due to overly complex probability density functions that fail to generalise.
- **Underfitting** occurs when a large smoothing alpha α leads to overly simple probability density functions, resulting in low training and validation accuracies.



- High Variance, Low Bias
- Complex Probability Density
- High Train Accuracy
- Low Validation Accuracy





- High Bias, Low Variance
- Simple Probability Density
- Low Train Accuracy
- Low Validation Accuracy

Why Unigram ≠ Enough for Sentiment or Topic Nuance

A pure bag-of-words (unigrams) treats each token independently, so "good" and "not good" contribute almost the same evidence. Likewise, "important decision" vs "unimportant decision" share the token *decision*.

What Plain Unigrams Miss	How Small *N-Grams* Fix It	Practical Recipe (scikit-learn)
Negation flips polarity "good" vs "not good"; "like" vs "don't like".	The bigram *not good* becomes its own feature, giving the classifier a separate weight from *good*.	CountVectorizer(ngram_range=(1,2)) keeps unigrams and bigrams; set min_df≥2 to drop ultra-rare pairs.
Intensifiers change magnitudeplain *good* vs **"very good", "so bad", "absolutely fantastic"*.	Bigram or trigram captures *very good*, *so bad* etc., allowing a larger positive/negative weight than the base adjective.	Keep intensifiers in the stop-word list; or pre-combine: regex `very\ (good

Example: *"I do not really like this movie at all."*

Feature Type	Tokens Extracted
Unigrams (1-grams)	i, do, not, really, like, this, movie, at, all
Bigrams (2-grams)	i do, do not, not really, really like, like this, this movie, movie at, at all

TF-IDF (Term Frequency-Inverse Document Frequency)

Term Frequency (TF): Measures how often a word appears in a document. A higher frequency suggests greater importance. If a term appears frequently in a document, it is likely relevant to the document's content.

$$TF(t,d) = \frac{\text{Number of times term } t \text{ appears in document } d}{\text{Total number of terms in document } d}$$

Common Locally

Inverse Document Frequency (IDF): Reduces the weight of common words across multiple documents while increasing the weight of rare words. If a term appears in fewer documents, it is more likely to be meaningful and specific.

$$IDF(t,d) = \log \frac{\text{Total number of documents in corpus } D}{\text{Number of documents containing term } t}$$

Rare Globally

Unlike simple word frequency (Bag-of-Words), *TF-IDF* balances common and rare words to highlight the most meaningful terms.

TF-IDF: Example

#1: The cat sat on the mat.
#2: The dog played in the park.
#3: The cat and dog are both great.

The word "cat" appear 1 time. The total of terms in document #1 is 6.

$\rightarrow \mathrm{TF}(\text{"cat"}, \#1) = \frac{1}{6}.$

The total number of documents in the corpus (D) is 3. The number of documents containing the term "cat" is 2.

$\rightarrow IDF(\text{"cat"}, D) = \log\frac{3}{2} = 0.405.$

$$\mathrm{TF{-}IDF}(\text{"cat"}, \#1, D) = \frac{1}{6} \times \log\frac{3}{2} = 0.068.$$

TF-IDF: Example

#1: The cat sat on the mat.
#2: The dog played in the park.
#3: The cat and dog are both great.

Unique Words in Corpus

$$\text{Vocabulary} = \{\text{and, are, both, cat, dog, great, in, mat, on, park, played, sat, the}\}$$

		and	are	both	cat	dog	great	in	mat	on	park	played	sat	the
The cat sat on the mat.	TF	$\frac{0}{6}$	$\frac{0}{6}$	$\frac{0}{6}$	$\frac{1}{6}$	$\frac{0}{6}$	$\frac{0}{6}$	$\frac{0}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{0}{6}$	$\frac{0}{6}$	$\frac{1}{6}$	$\frac{2}{6}$
	IDF	$\log\frac{3}{1}$	$\log\frac{3}{1}$	$\log\frac{3}{1}$	$\log\frac{3}{2}$	$\log\frac{3}{2}$	$\log\frac{3}{1}$	$\log\frac{3}{1}$	$\log\frac{3}{1}$	$\log\frac{3}{1}$	$\log\frac{3}{1}$	$\log\frac{3}{1}$	$\log\frac{3}{1}$	$\log\frac{3}{3}$
	TF-IDF	0	0	0	$\frac{1}{6}\times\log\frac{3}{2}$	0	0	0	$\frac{1}{6}\times\log\frac{3}{1}$	$\frac{1}{6}\times\log\frac{3}{1}$	0	0	$\frac{1}{6}\times\log\frac{3}{1}$	$\frac{2}{6}\times\log\frac{3}{3}$

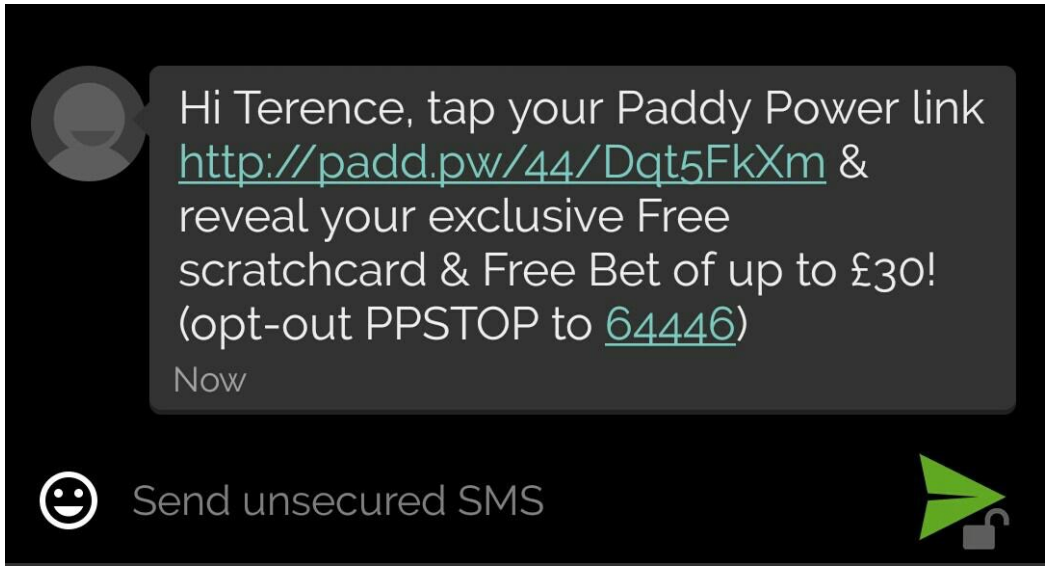
(2/6)x0 = 0

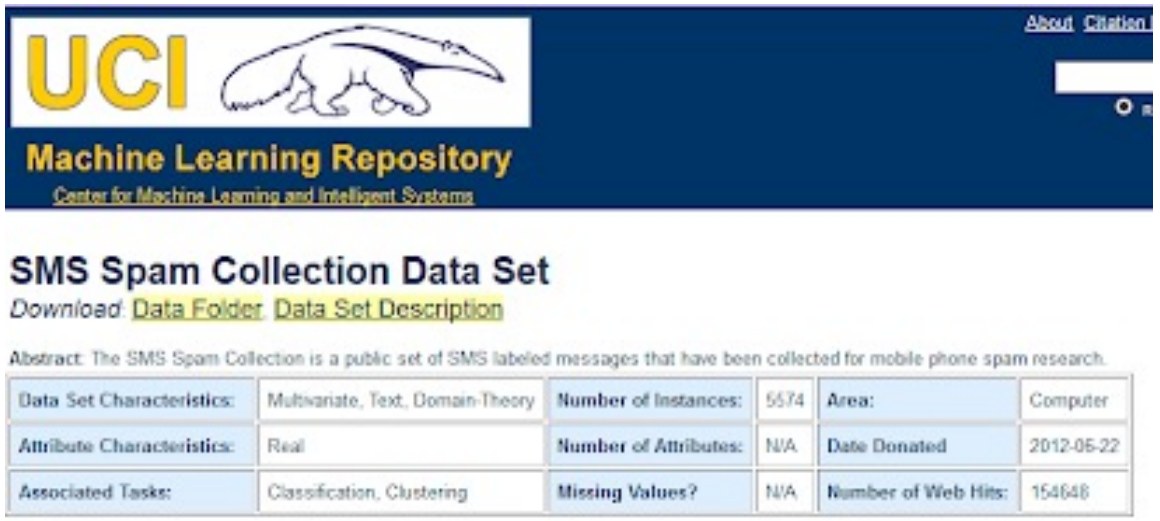
Q: With TF-IDF, will it make a difference if we are to remove stop words or not?

SMS Spam Dataset

5,574 SMS Messages

- 2 Classes, *Spam* and *Ham*
- 7,451 Vocabulary Size





`https://archive.ics.uci.edu/dataset/228/sms+spam+collection`

Bag-of-Words vs TF-IDF

Model + Features	F1 Score	What the score tells us	SMS-Specific Rationale
MNB + TF (α = 1)	0.949	NB excels when a *single* "spam clue" (free, win, cash) should immediately trigger the spam label. Raw counts let those clues contribute their full log-likelihood.	• Messages are **very short** (≈ 15 tokens); a spam word often appears **once** → NB's probability jump is decisive. • Spam vocabulary is **repetitive and class-specific**; IDF would unnecessarily dampen those high-frequency spam tokens.
MNB + TF-IDF (α = 0.1)	0.944	Still strong, but IDF down-weights spam words that appear in *many* spam messages, so NB's evidence is slightly muted.	• Words like "call" and "now" are common to both classes; TF-IDF helps by shrinking their weight – yet it *also* shrinks genuinely useful spam cues that aren't globally rare, costing a little recall/precision.
LR + TF-IDF (C = 1000)	0.939	LR benefits from TF-IDF because it reduces the dominance of background words; high C shows the model wanted minimal regularisation once features were re-weighted.	• With TF-IDF each SMS vector is length-normalised → LR no longer biases toward longer ham texts. • Still trails NB because LR needs more data to learn strong weights for every spam word; NB's generative counts capture that with fewer examples.
LR + TF (C = 1000)	0.936	Lowest score: raw counts make very common tokens ("call", "now") huge; LR can only counter by giving those words tiny or negative weights, a harder optimisation problem in this tiny dataset.	• Sparse, high-dimensional space + few spam samples ⇒ LR risks over- or under-weighting features despite the high C. • Without IDF, long ham messages get larger feature norms, nudging LR toward ham predictions.

Summary

- Naïve Bayes (NB) applies Bayes' theorem under the conditional-independence assumption, turning joint likelihoods into simple 1-D products. Training reduces to counting frequencies or estimating a mean + variance, so NB scales to tens-of-thousands of features in seconds and delivers well-calibrated posterior probabilities.
- Multinomial Naïve Bayes (MNB) interprets every token count as independent evidence; α-smoothing guards against zero probabilities. Because a single rare spam word can flip the posterior, Multinomial NB often beats heavier models on short, sparse documents (spam, sentiment, topic tagging) and needs little data to do so.
- Gaussian Naïve Bayes (GNB) assumes each feature is normally distributed per class, storing only μ and σ^2. It handles real-valued inputs instantly, yet the Gaussian fit is sensitive to outliers—even a few extreme points can skew means/variances and degrade accuracy, so robust preprocessing is essential.
- Both MNB and GNB train ~100× faster than discriminative models, e.g. Logistic Regression, and work well when samples are scarce or features far outnumber observations. Word- or feature-likelihood tables are directly interpretable and can highlight the strongest signals for each class.
- MNB' strong multinomial assumptions yield high bias but low variance: it may underfit when features are correlated or not truly multinomial, yet rarely overfits—even on small data. Proper Laplace (α) smoothing is key: too small α overemphasizes rare counts (risking overfitting), while too large α or heavy smoothing produces a bland, underfitted model.