Bayes Rule and Bayes Classifier

Liang Liang

Classification and Bayes Rule

- We can use Bayes Rule for classification.

- We will use simple examples to show how to apply Bayes decision rule for Binary Classification (two classes)

- Bayes rule can be applied for multiclass classification.

- Bayes decision rule will help us explain/understand classification using the language of probability and statistics.

Binary Classification: two classes

- We have a set of data points $\{x_1, x_2, x_3, \dots, x_N\}$ and $x_n \in \mathcal{R}^M$
- We have a set of labels $\{y_1, y_2, y_3, \dots, y_N\}$ and $y_n \in \{0, 1\}$
- The data points are from two classes. y_n is the class label of x_n
- One data point belongs to only one class
- Each class has a probability density function (PDF), from which the data points are "generated"
  - $p(x|y = 0)$ is the PDF of class-0, x refers to a data point, y is the label
  - $p(x|y = 1)$ is the PDF of class-1
- Each class has a prior probability: $\pi_0 = p(y = 0)$ and $\pi_1 = p(y = 1)$

Example: each class has a M-D Gaussian PDF

- Each class has a M-D Gaussian PDF

PDF of class-0: $p(x|y=0) = \mathcal{N}(x; \mu_0, \Sigma_0)$, parameters: μ_0 and Σ_0

PDF of class-1: $p(x|y=1) = \mathcal{N}(x; \mu_1, \Sigma_1)$, parameters: μ_1 and Σ_1

- Each class has a prior probability: $\pi_0 = p(y=0)$ and $\pi_1 = p(y=1)$

$$\pi_0 + \pi_1 = 1$$

Example: each class has a 1-D Gaussian PDF

- Each class has a 1D Gaussian PDF

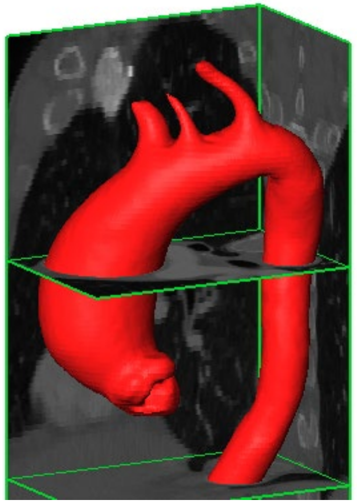
$$p(x|y=0) = \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{(x-\mu_0)^2}{2\sigma_0^2}}$$, parameters: μ_0 and σ_0

$$p(x|y=1) = \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}$$, parameters: μ_1 and σ_1

- Each class has a prior probability: $\pi_0 = p(y=0)$ and $\pi_1 = p(y=1)$

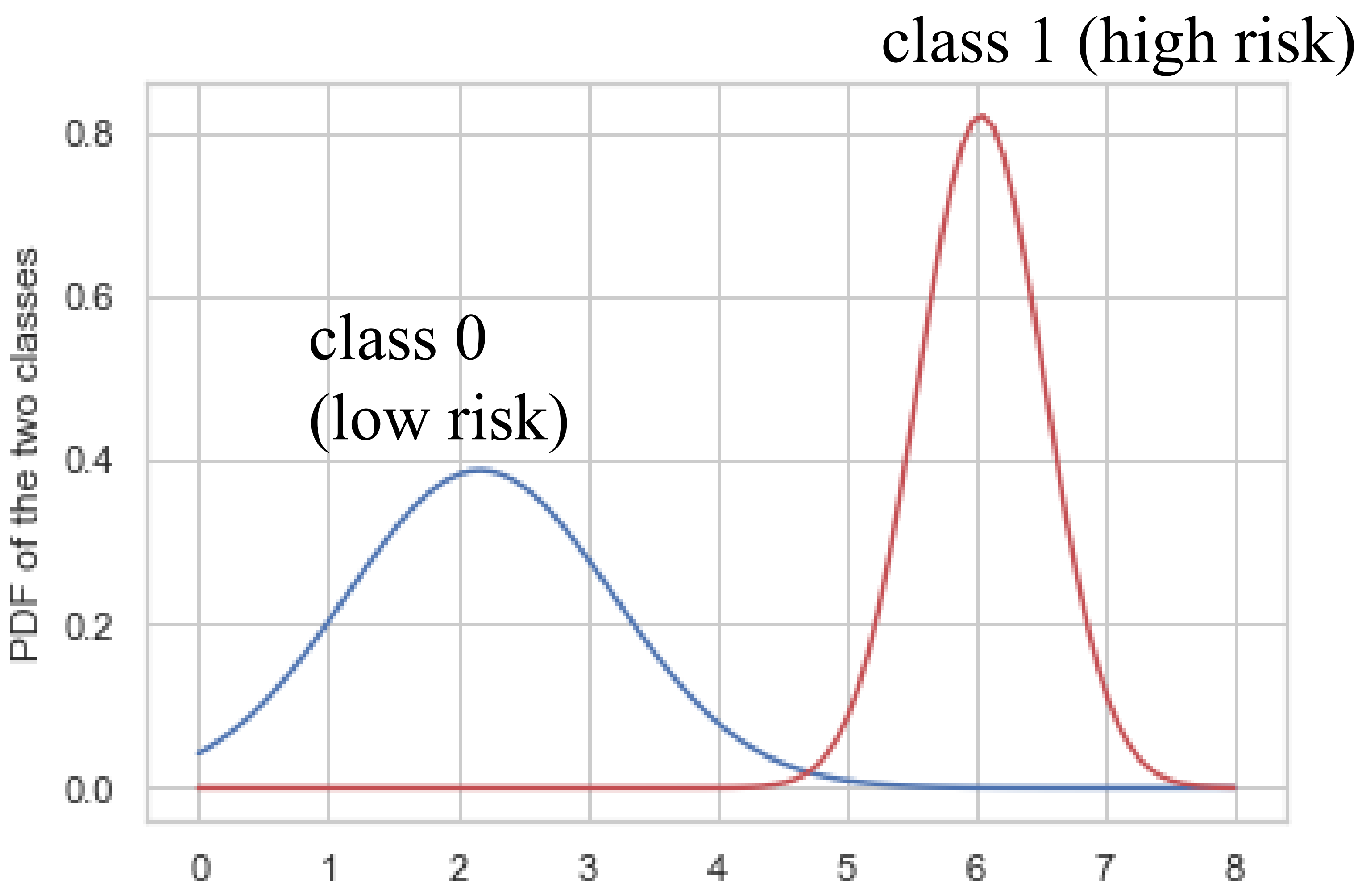
$$\pi_0 + \pi_1 = 1$$

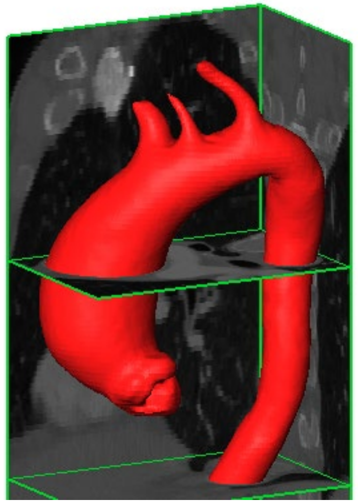
Example: aortic aneurysm



- Millions of people have aortic aneurysms.
- If an aortic aneurysm of a patient ruptures, the patient will die very soon if left untreated.
- The rupture risk is measured by using aneurysm diameter.
- x_n is aneurysm diameter of a patient (indexed by n).
- The current clinical decision rule:

  perform surgery (cut aneurysm) if $x_n > t$
- To find the best threshold t, assume we can collect some data:
  - (class-0) aneurysm diameters of the patients whose aneurysms did <u>not rupture</u> in the last ten years.
  - (class-1) aneurysm diameters of the patients whose aneurysms did rupture in the last ten years.





x_n is the aortic aneurysm diameter of a patient.

Decision Rule: whether to cut the aneurysm or not, based on a threshold **t**

if x > **t**, then x belongs to class 1 (cut)

if x < **t**, then x belongs to class 0 (not cut).

What is the “optimal” value of **t** ?

Notations

X is a random variable (vector), and Y is a random variable (class label)

x is a data point – an observation of X

y is a class label – an observation of Y (0 or 1 for binary classification)

For simplicity:

we use $p(x|y)$ to represent $p(X = x|Y = y)$

we use $p(y|x)$ to represent $p(Y = y|X = x)$

Notations

x is a data point

y is a class label (0 or 1 for binary classification)

$\boldsymbol{p}(\boldsymbol{x}|\boldsymbol{y})$ is the prob. density function (PDF) of the class y (class-y PDF)

- the distribution of the data points (e.g., x) in a given class (e.g., y)

$\boldsymbol{p}(\boldsymbol{y}|\boldsymbol{x})$ is the posterior probability that is used for classification

- prob. of a data point x belonging to the class y
- if $p(y = 0|x) > (y = 1|x)$ then x is classified into the class 0

Bayes Rule/Theorem

class-y PDF value at the point x

class prior

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

posterior: the probability of the data point x belonging to class-y

probability density of the data, known as the normalization constant

Law of total probability

$$p(x) = \sum_{k=1}^{K} p(y = k)p(x|y = k)$$

probability density of the data from all the classes

prior probability of the class y, where y= k

probability density of the data in one single class y, where y=k

Bayes Rule: when $p(x|y) = \mathcal{N}(x; \mu_y, \Sigma_y)$, $p(y) = \pi_y$

class-y PDF value at the point x

class prior

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} = \frac{\mathcal{N}(x; \mu_y, \Sigma_y)\pi_y}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x; \mu_k, \Sigma_k)}$$

posterior: the probability of the data point x belonging to class-y

prob. density of the data

K=2 for binary classification

$$p(x) = \sum_{k=1}^{K} p(y = k)p(x|y = k) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x; \mu_k, \Sigma_k)$$

Bayes Rule

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

for binary classification

- compute $p(y)$: $p(y = 0)$ and $p(y = 1)$

$$p(y = 1)$$

What is the probability that an aneurysm ruptures $(y = 1)$ in a patient ?
note: we do not have any diameter measurement of the aneurysm

Bayes Rule

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

for binary classification

- compute $p(y)$: $p(y = 0)$ and $p(y = 1)$
- compute $p(x|y)$: $p(x|y = 0)$ and $p(x|y = 1)$

$$p(x|y = 0)$$

compute the probability density at the data point x using the PDF of the class 0

given $y = 0$: under the condition that the class label is *0*

Bayes Rule

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

for binary classification

- compute $p(y)$: $p(y = 0)$ and $p(y = 1)$
- compute $p(x|y)$: $p(x|y = 0)$ and $p(x|y = 1)$
- compute $p(x)$: use law of total probability

$$p(x) = p(y = 0)p(x|y = 0) + p(y = 1)p(x|y = 1)$$

Bayes Rule

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

for binary classification

- compute $p(y)$: $p(y=0)$ and $p(y=1)$
- compute $p(x|y)$: $p(x|y=0)$ and $p(x|y=1)$
- compute $p(x)$: use law of total probability
- compute $p(y|x)$: $p(y=0|x)$ and $p(y=1|x)$

$$p(y=0|x)$$

y is the class label of x

compute the probability that the class label of x is 0

given x: we have obtained a data point x

Bayes Classifier for binary classification

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} = \frac{\mathcal{N}\left(x; \mu_y, \Sigma_y\right)\pi_y}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x; \mu_k, \Sigma_k)}$$

- Classification Rule of a Bayes Binary Classifier
  - if $p(y = 0|x) > (y = 1|x)$ then x is classified into the class 0
  - if $p(y = 0|x) < (y = 1|x)$ then x is classified into the class 1
  - if $p(y = 0|x) = (y = 1|x)$ then x is on the decision boundary
- The internal parameters of $p(y|x)$ will be determined using training data

Training/Learning a Bayes classifier for binary classification (fit model to data to get the best parameters of the PDFs)

- We have a set of <u>training</u> data points $\{x_1, x_2, x_3, \dots, x_N\}$ and $x_n \in \mathcal{R}^M$
- We have <u>a set of class labels</u> $\{y_1, y_2, y_3, \dots, y_N\}$ and $y_n \in \{0, 1\}$
- Assume the data points are observations of i.i.d. random variables
- The optimal parameters of the classifier can be obtained by minimizing negative log likelihood loss (NLL)

$$
\begin{aligned}
loss &= -log\left(\prod_{n=1}^{N} p(x_n, y_n)\right) \\
&= -\sum_{n=1}^{N} log(p(x_n, y_n)) \\
&= -\sum_{n=1}^{N} log(p(y_n)p(x_n|y_n))
\end{aligned}
$$

textbook: Machine-Learning-A-Probabilistic-Perspective

Training/Learning the Bayes classifier (Assuming $p(x|y) = \mathcal{N}(x; \mu_y, \Sigma_y)$, $p(y) = \pi_y$)

- Assuming class PDF $p(x|y) = \mathcal{N}(x; \mu_y, \Sigma_y)$, then

$$p(x_n|y_n) = \mathcal{N}(x_n; \mu_{y_n}, \Sigma_{y_n})$$

- Minimize the NLL loss to get the best values of π_y, μ_y and Σ_y

$$loss = -\sum_{n=1}^{N} log\left(\pi_{y_n}\mathcal{N}(x_n; \mu_{y_n}, \Sigma_{y_n})\right)$$

The solution:

$$\pi_y = \frac{the\ number\ of\ data\ points\ in\ the\ class\ y}{N}$$

μ_y and Σ_y are calculated using the data points in class-y

textbook: Machine-Learning-A-Probabilistic-Perspective

Apply the Bayes classifier to classify a data point

- We have a new data point x that is not in the training dataset
- We may not know the class label y of x
- We apply the trained classifier to classify the data point x
    - Compute the posterior probability $p(y|x) = \frac{p(x|y)\,p(y)}{p(x)}$
    - Classification rule (binary classifiation):

$$\text{if } p(y = 0|x) > p(y = 1|x),\ \text{ then } x \text{ belongs to class-0}$$

$$\text{if } p(y = 0|x) < p(y = 1|x),\ \text{ then } x \text{ belongs to class-1}$$

$$\text{if } p(y = 0|x) = p(y = 1|x),\ \text{ then } x \text{ is on decision boundary}$$

textbook: Machine-Learning-A-Probabilistic-Perspective

Apply the Bayes classifier

$$p(y=0|x) = \frac{p(x|y=0)\,p(y=0)}{p(x)}$$

$$p(y=1|x) = \frac{p(x|y=1)\,p(y=1)}{p(x)}$$

- Make classification using log-likelihood ratio $h(x)$

$$h(x) = log\frac{p(y=0|x)}{p(y=1|x)} = log\frac{p(y=0)p(x|y=0)}{p(y=1)p(x|y=1)}$$

where $p(y=1|x)$ is assumed not equal to 0

if $h(x) > 0$, then then x belongs to class-0 ($p(y=0|x) > p(y=1|x)$)

if $h(x) < 0$, then then x belongs to class-1 ($p(y=0|x) < p(y=1|x)$)

Then we do not need to compute $p(x)$

The curve/surface defined by $h(x) = 0$ is the **decision boundary**

textbook: Machine-Learning-A-Probabilistic-Perspective

Bayes classifier - the decision boundary
(Assuming $p(x|y) = \mathcal{N}\left(x; \mu_y, \Sigma_y\right)$, $p(y) = \pi_y$)

- Classification using log-likelihood ratio $h(x)$

$$h(x) = log \frac{\pi_0 \mathcal{N}(x; \mu_0, \Sigma_0)}{\pi_1 \mathcal{N}(x; \mu_1, \Sigma_1)}$$

$h(x) = 0$ defines decision boundary (when $p(y = 0|x) = p(y = 1|x)$)

if $\Sigma_0 \neq \Sigma_1$, then $h(x) = 0 \Longrightarrow x^T A x - b^T x + c = 0$

a quadratic surface/curve

if $\Sigma_0 = \Sigma_1$, then $h(x) = 0 \Longrightarrow a^T x - d = 0$

hyperplane or a line

Bayes classifier: Accuracy and Error

- The accuracy/error of a Bayes classifier can be calculated using equations.

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

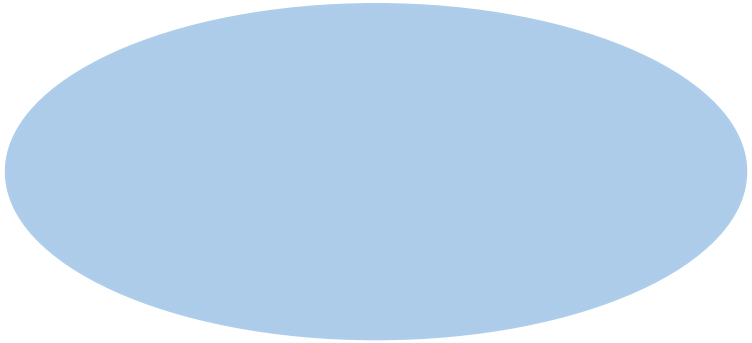
Notations

- $x \to C_y$ means a data point *x* is classified into the class *y* by using a Bayes classifier.
  - $x \to \boldsymbol{C_0}$: *x* is classified into the class *0* by using the classifier
  - $x \to \boldsymbol{C_1}$: *x* is classified into the class *1* by using the classifier

- $x \in C_y$ means a data point *x* is in the class *y*
  - $x \in \boldsymbol{C_0}$ means a data point *x* is in the class *0*
  - $x \in \boldsymbol{C_1}$ means a data point *x* is in the class *1*

A basic property of class PDFs

class 0

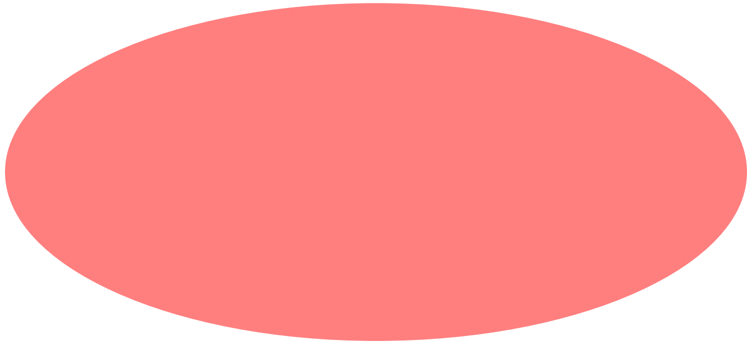


$\boldsymbol{p(x|y=0)}$ is the PDF of the class 0

$$\int_{x \in \boldsymbol{C_0}} \boldsymbol{p(x|y=0)}\,dx = 1$$

A basic property of class PDFs

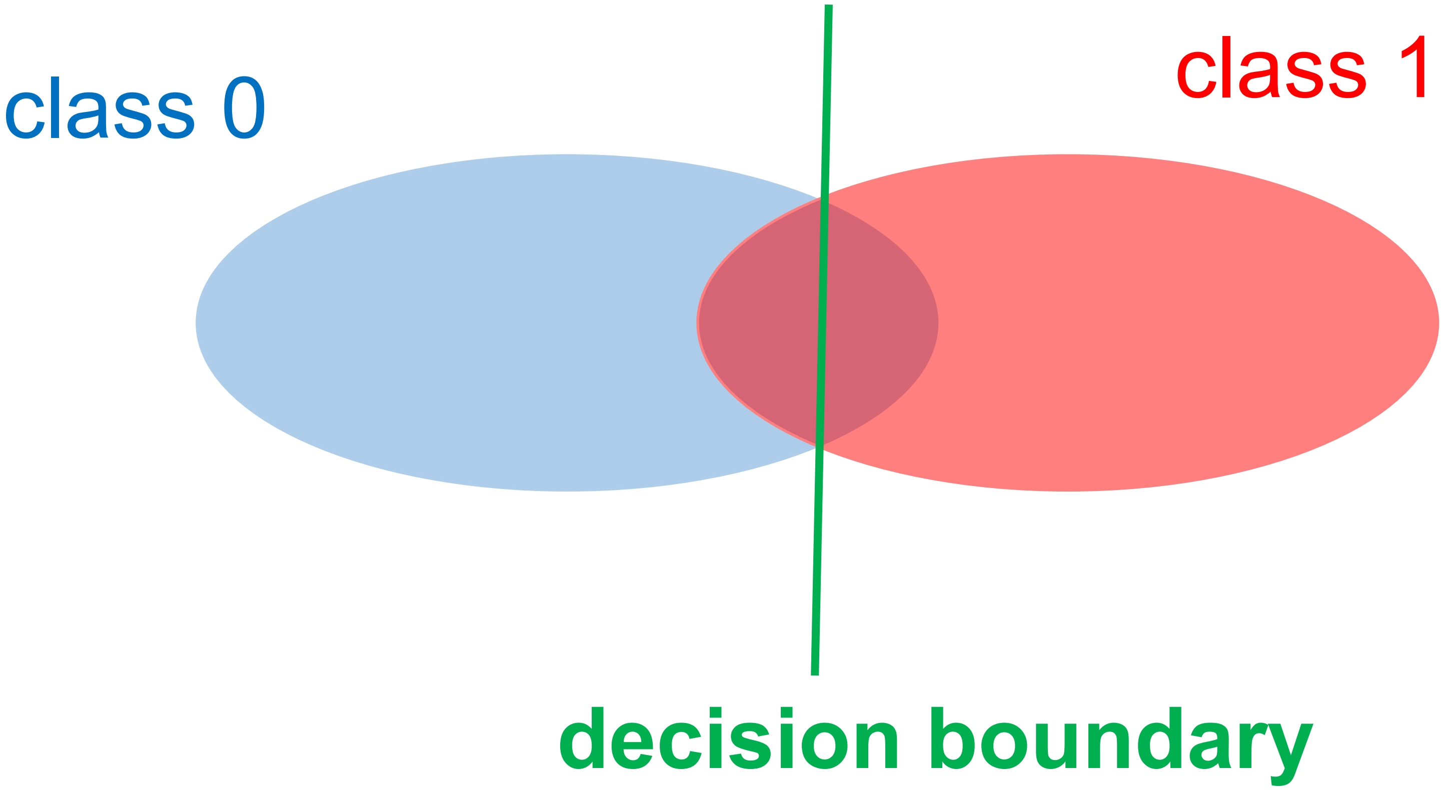
class 1



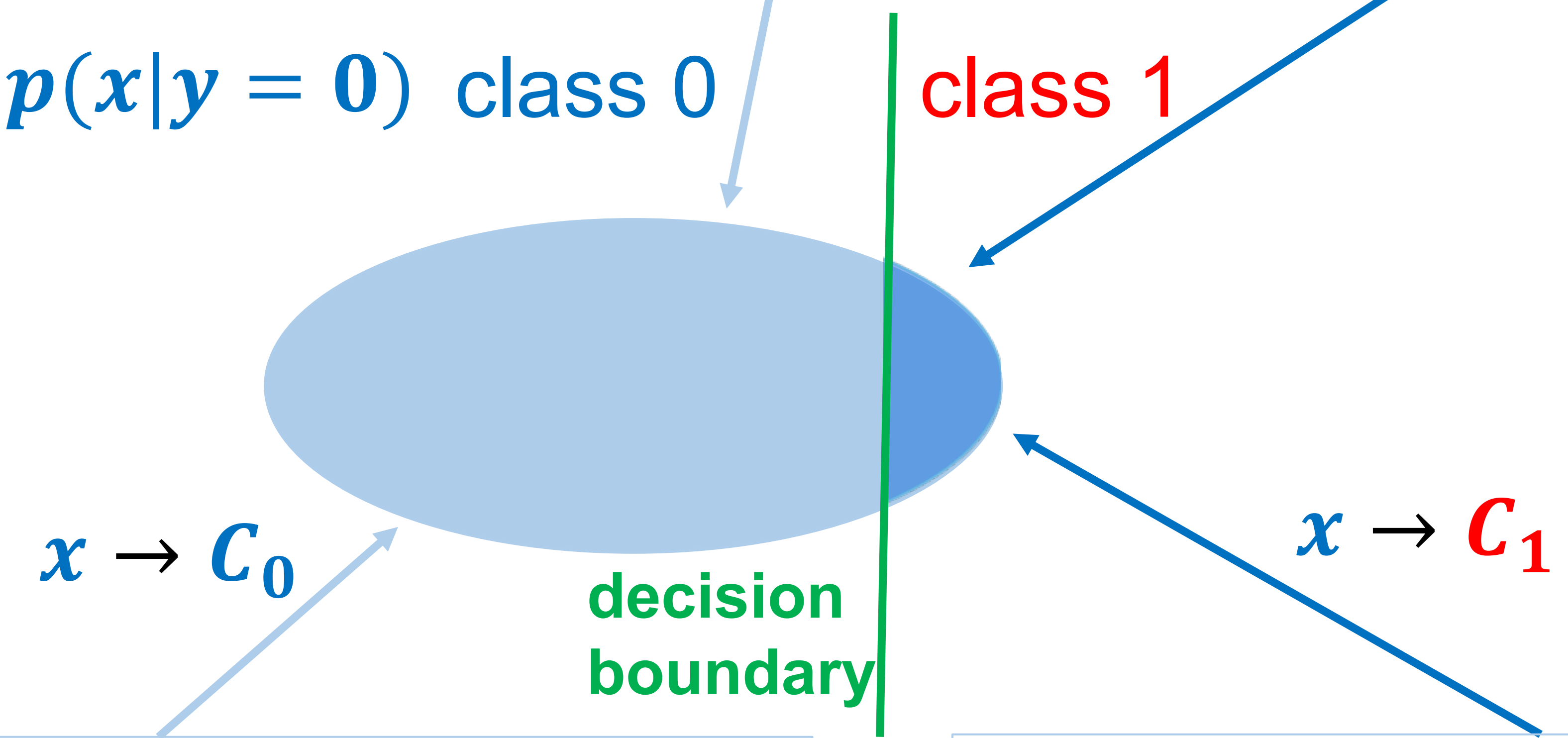
$\boldsymbol{p(x|y=1)}$ is the PDF of the class 1

$$\int_{x\in \boldsymbol{C_1}} \boldsymbol{p(x|y=1)}\,dx = 1$$

binary classification



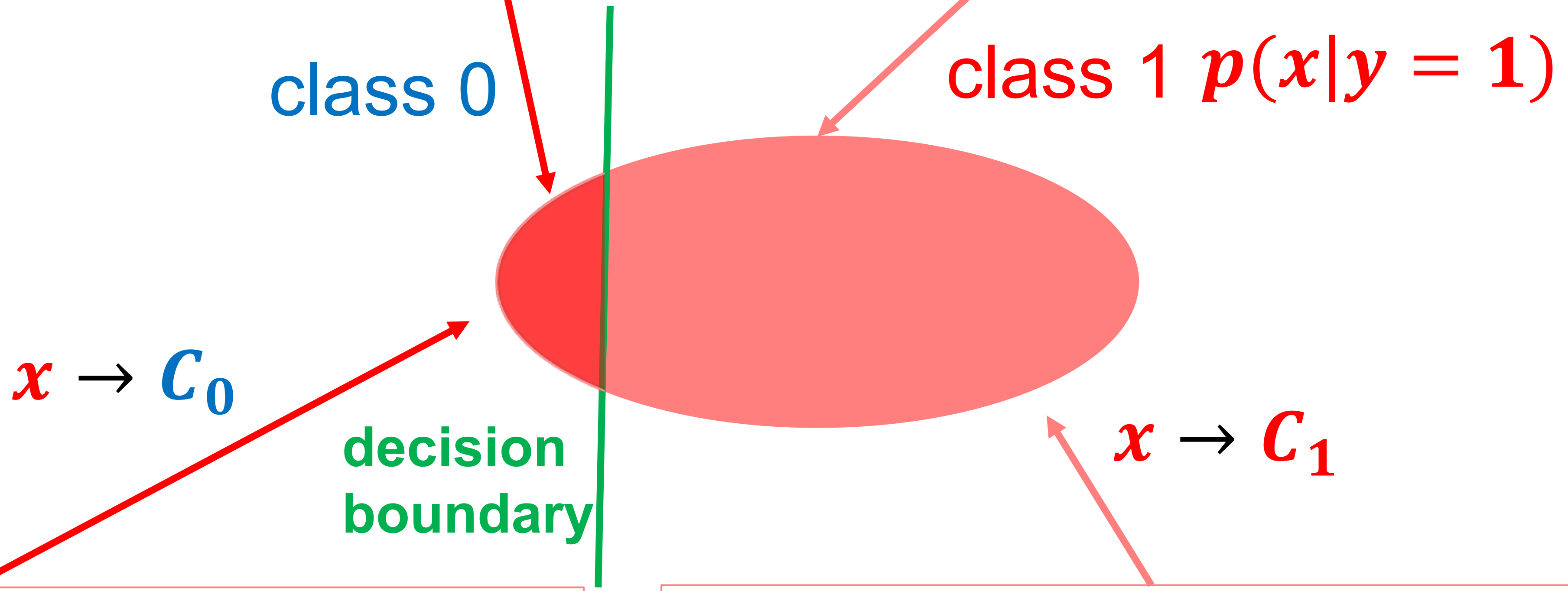
$$1 = \int_{x \in C_0} p(x|y=0)dx = \int_{x \to C_0} p(x|y=0)dx + \int_{x \to C_1} p(x|y=0)dx$$



right classifications for class 0: data points with true label=0 are classified into class 0

wrong classifications for class 0: data points with true label=0 are classified into class 1

$$1 = \int_{x \in C_1} p(x|y=1)dx = \int_{x \to C_0} p(x|y=1)dx + \int_{x \to C_1} p(x|y=1)dx$$



wrong classifications for class 1: data points with true label=1 are classified into class 0

right classifications for class 1: data points with true label=1 are classified into class 1

binary classification accuracy

$$accuracy = \int_{x\in\Omega_{correct}} p(x)dx$$

$$\Omega_{correct} = \{x|x\ is\ correctedly\ classified\}$$

$x\ is\ correctedly\ classified$: then (1) or (2) is true

(1) true label $y = 0$ and $x \to C_0$

(2) true label $y = 1$ and $x \to C_1$

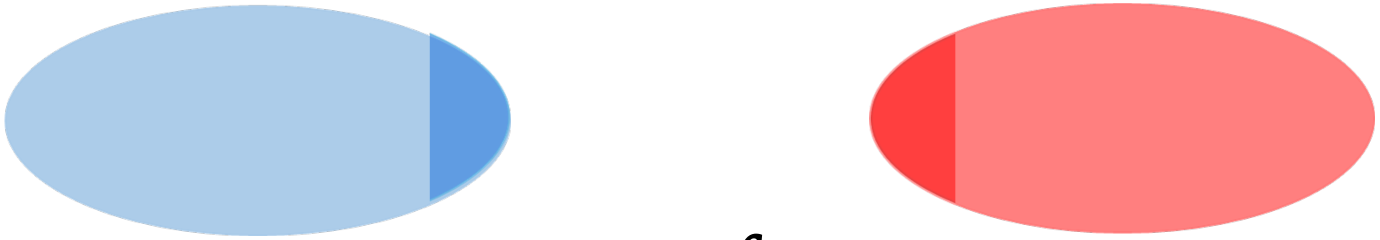
$\Omega_{correct} = \{x|y = 0, x \to C_0\} \cup \{x|y = 1, x \to C_1\}$

binary classification accuracy

$$accuracy = \int_{\boldsymbol{x}\in\boldsymbol{\Omega}_{\boldsymbol{correct}}} p(x)dx$$

$$= \int_{\boldsymbol{x}\in\boldsymbol{\Omega}_{\boldsymbol{correct}}} [(p(y=0)p(x|y=0) + p(y=1)p(x|y=1))]dx$$

$$= \int_{\boldsymbol{x}\in\boldsymbol{\Omega}_{\boldsymbol{correct}}} p(y=0)p(x|y=0)dx + \int_{\boldsymbol{x}\in\boldsymbol{\Omega}_{\boldsymbol{correct}}} p(y=1)p(x|y=1)dx$$



$$= \int_{x\to C_0} p(y=0)p(x|y=0)dx + \int_{x\to C_1} p(y=1)p(x|y=1)dx$$

$\pi_0 = p(y=0)$ and $\pi_1 = p(y=1)$

binary classification accuracy

$$accuracy = \pi_0 \int_{\underbrace{x \to C_0}_{\substack{x \text{ is classified to class-0} \\ \text{(right classification)}}}} \overbrace{p(x|y=0)}^{\substack{x \text{ in class-0} \\ \text{(truth)}}} dx + \pi_1 \int_{\underbrace{x \to C_1}_{\substack{x \text{ is classified to class-1} \\ \text{(right classification)}}}} \overbrace{p(x|y=1)}^{\substack{x \text{ in class-1} \\ \text{(truth)}}} dx$$

Error rate: the expected classification error

$$error\ rate = 1 - accuracy$$

Error rate: the expected classification error

$$error\ rate = 1 - accuracy$$

$$= 1 - \left(\pi_0 \int_{x \to C_0} \boldsymbol{p(x|y=0)}dx + \pi_1 \int_{x \to C_1} \boldsymbol{p(x|y=1)}dx\right)$$

$$= (\pi_0 + \pi_1) - \pi_0 \int_{x \to C_0} \boldsymbol{p(x|y=0)}dx - \pi_1 \int_{x \to C_1} \boldsymbol{p(x|y=1)}dx$$

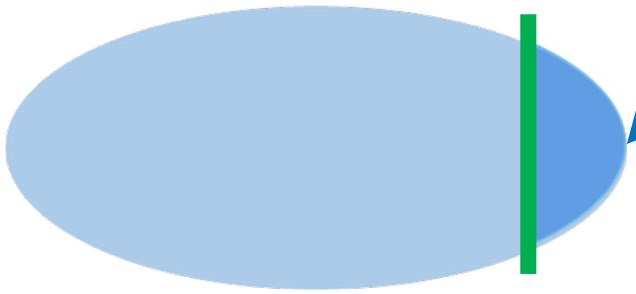
$$= \pi_0 \left(1 - \int_{x \to C_0} \boldsymbol{p(x|y=0)}dx\right) + \pi_1 \left(1 - \int_{x \to C_1} \boldsymbol{p(x|y=1)}dx\right)$$

Error rate: the expected classification error

$$error\ rate = 1 - accuracy$$

$$= \pi_0 \left(1 - \int_{x \to C_0} \boldsymbol{p(x|y=0)}dx \right) + \pi_1 \left(1 - \int_{x \to C_1} \boldsymbol{p(x|y=1)}dx \right)$$

$$\int_{x \to C_0} \boldsymbol{p(x|y=0)}dx + \int_{x \to C_1} \boldsymbol{p(x|y=0)}dx = 1$$

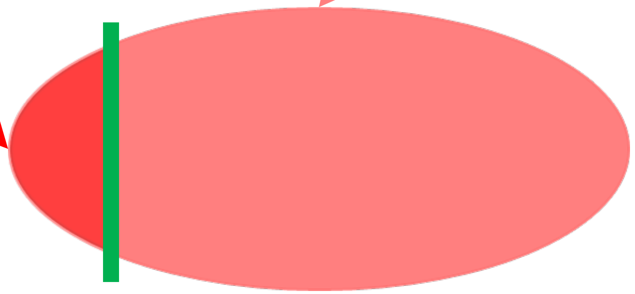


Error rate: the expected classification error

$$error\ rate = 1 - accuracy$$

$$= \pi_0 \left(1 - \int_{x \to \boldsymbol{C_0}} \boldsymbol{p(x|y=0)} dx \right) + \pi_1 \left(1 - \int_{x \to C_1} \boldsymbol{p(x|y=1)} dx \right)$$

$$\int_{x \to \boldsymbol{C_0}} \boldsymbol{p(x|y=1)} dx + \int_{x \to C_1} \boldsymbol{p(x|y=1)} dx = 1$$



Error rate: the expected classification error

$$error\ rate = 1 - accuracy$$

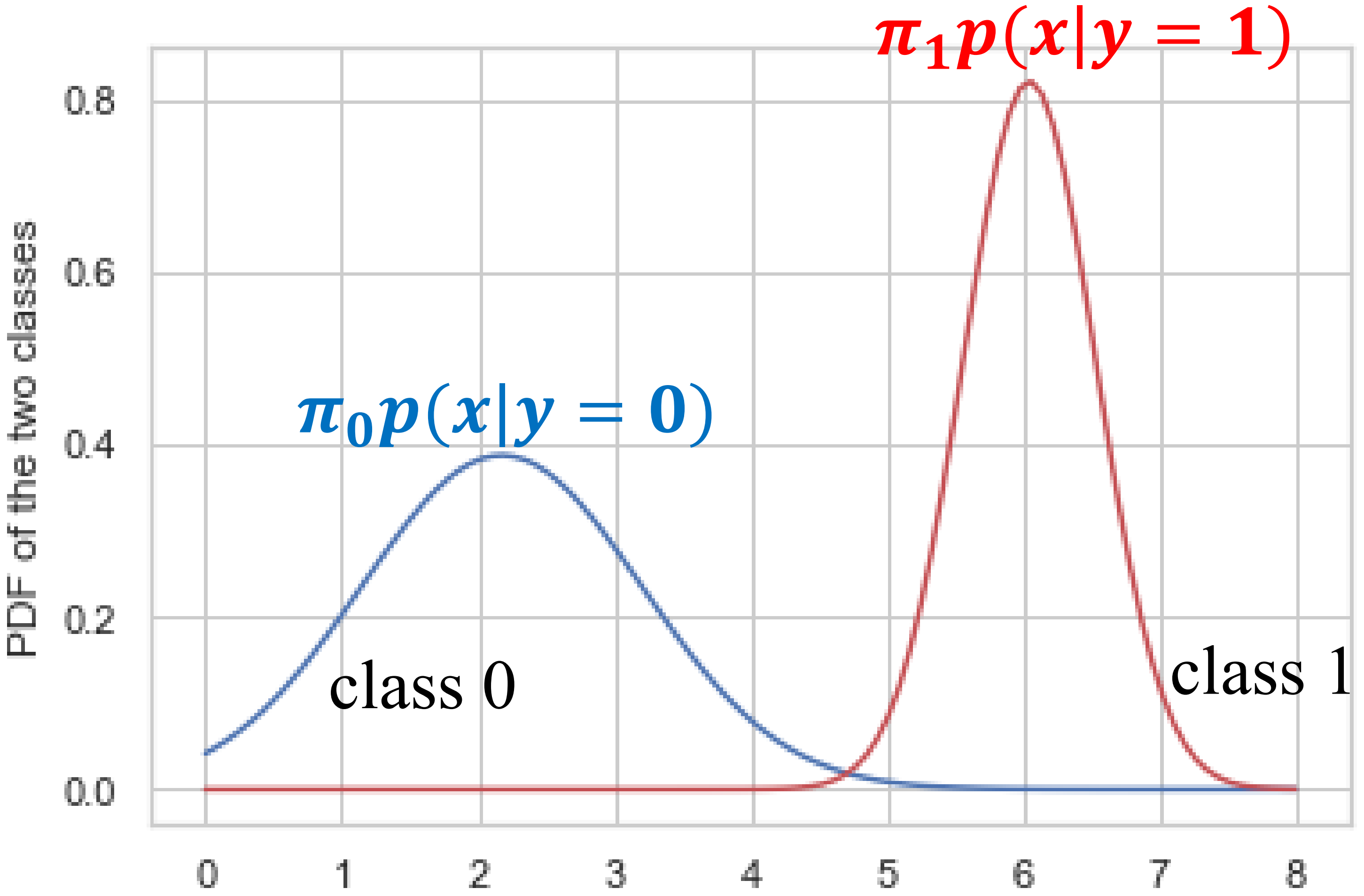
x in class-0 (truth)

x in class-1 (truth)

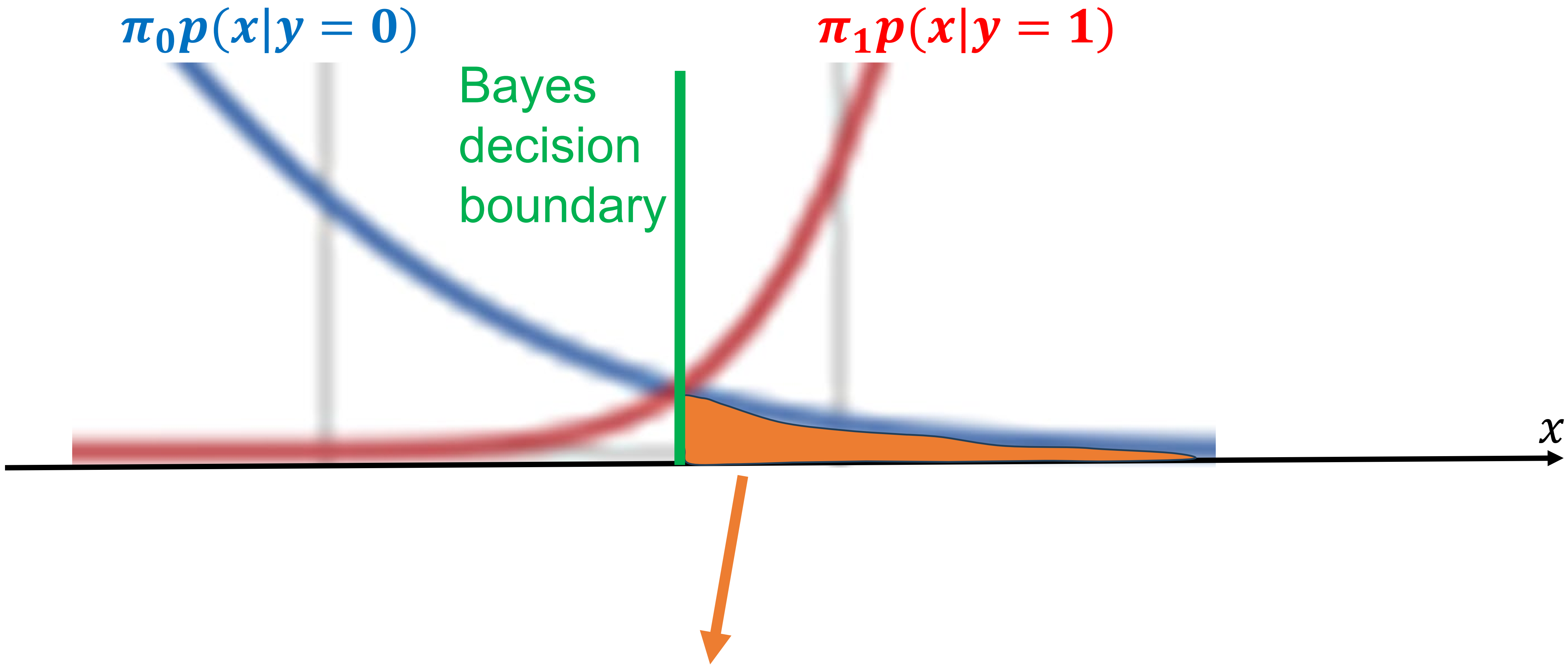
$$error\ rate = \pi_0 \int_{x \to C_1} \boldsymbol{p(x|y=0)}\,dx + \pi_1 \int_{x \to C_0} \boldsymbol{p(x|y=1)}\,dx$$

x is classified to class-1 (wrong classification)

x is classified to class-0 (wrong classification)



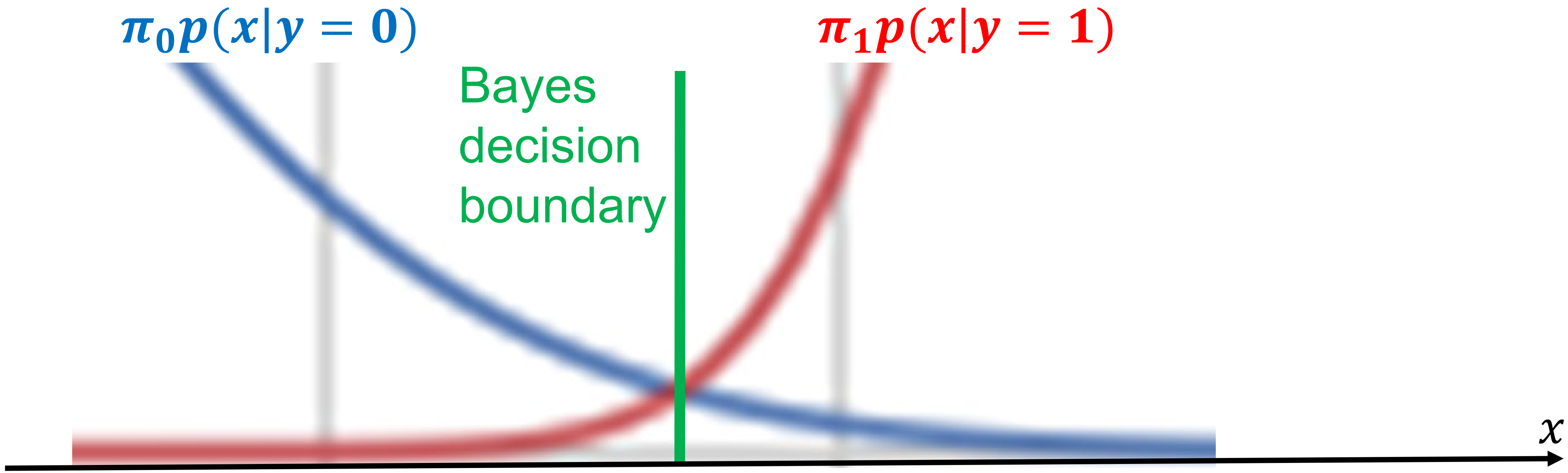
$\pi_1 p(x|y = 1)$
$\pi_0 p(x|y = 0)$
class 0
class 1
PDF of the two classes
0.0
0.2
0.4
0.6
0.8
0
1
2
3
4
5
6
7
8



What is the meaning of the **area** of the region in the orange color ?

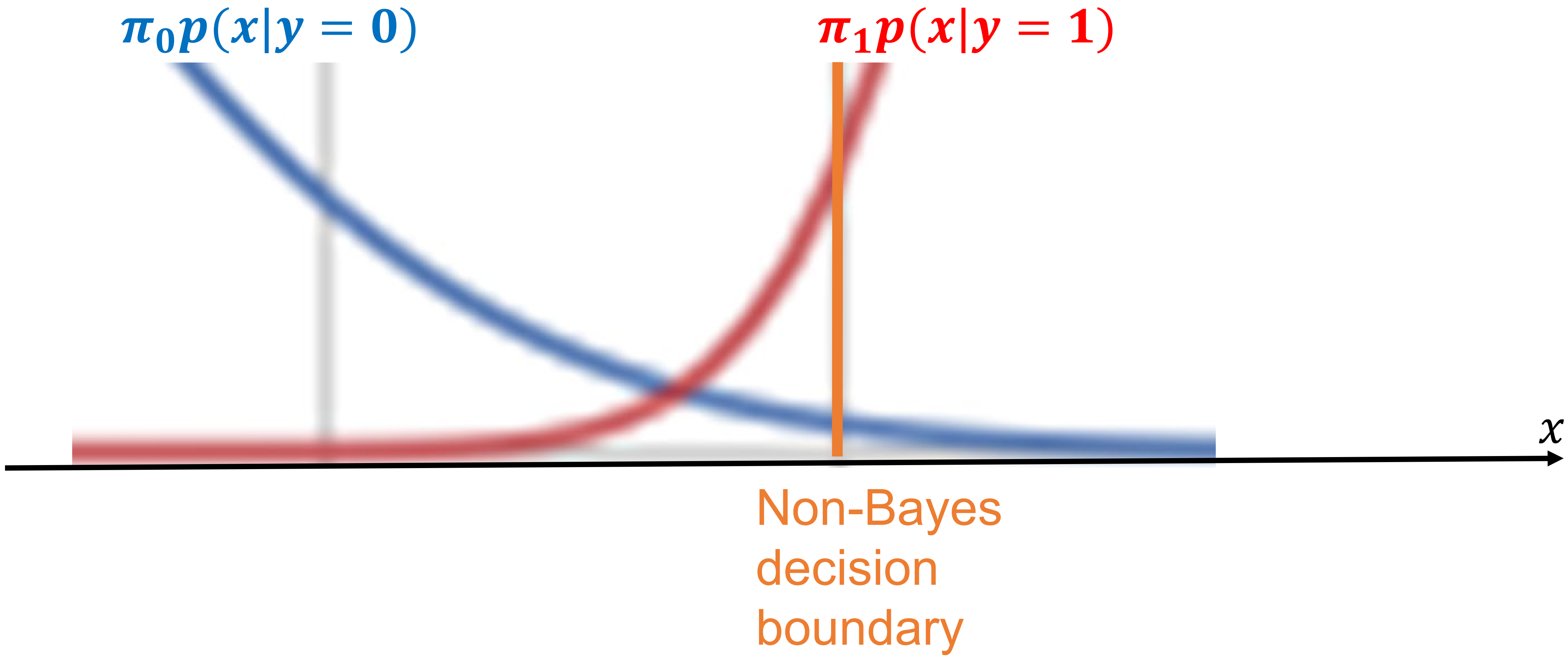
Area under the blue curve and on the right side of the decision boundary

(e.g., what does this mean: area=0.01 ?)

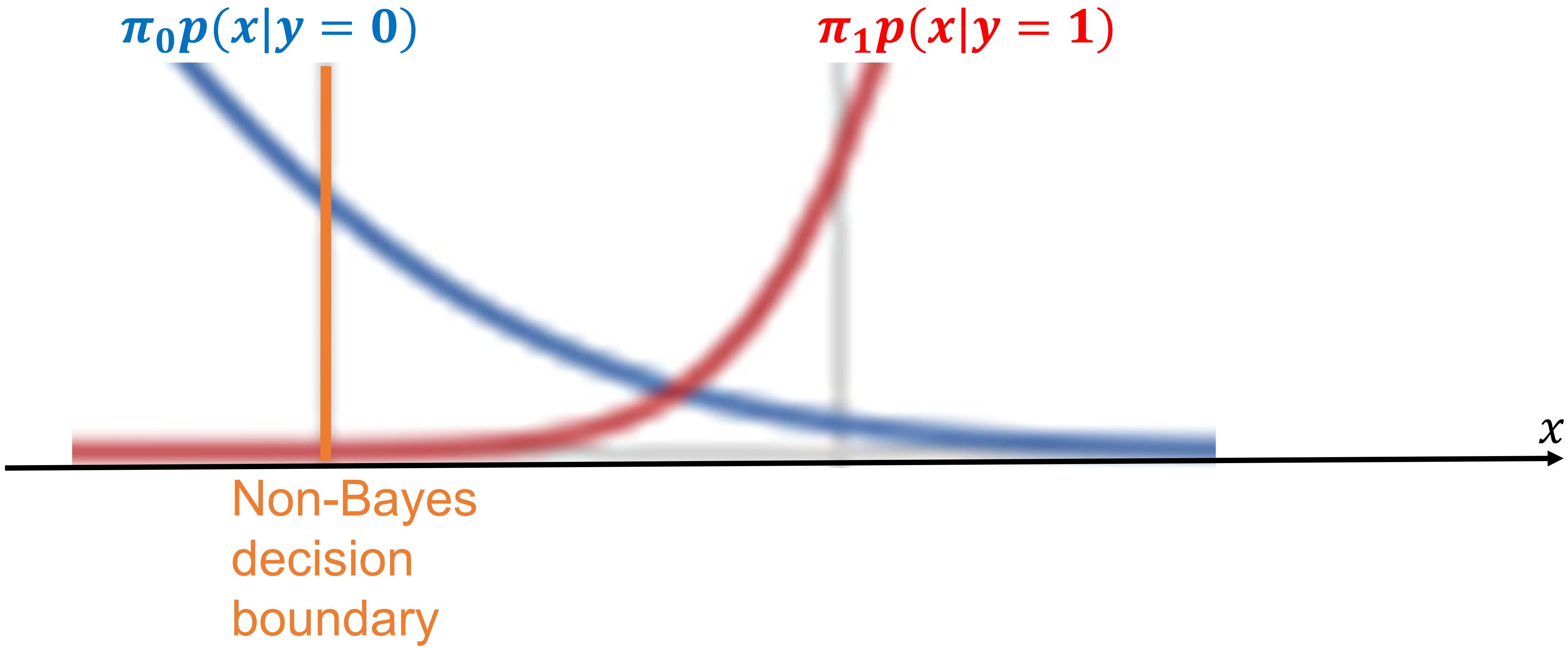


$$error\ rate = \pi_1 \int_{x \to C_0} p(x|y=1)dx + \pi_0 \int_{x \to C_1} p(x|y=0)dx$$

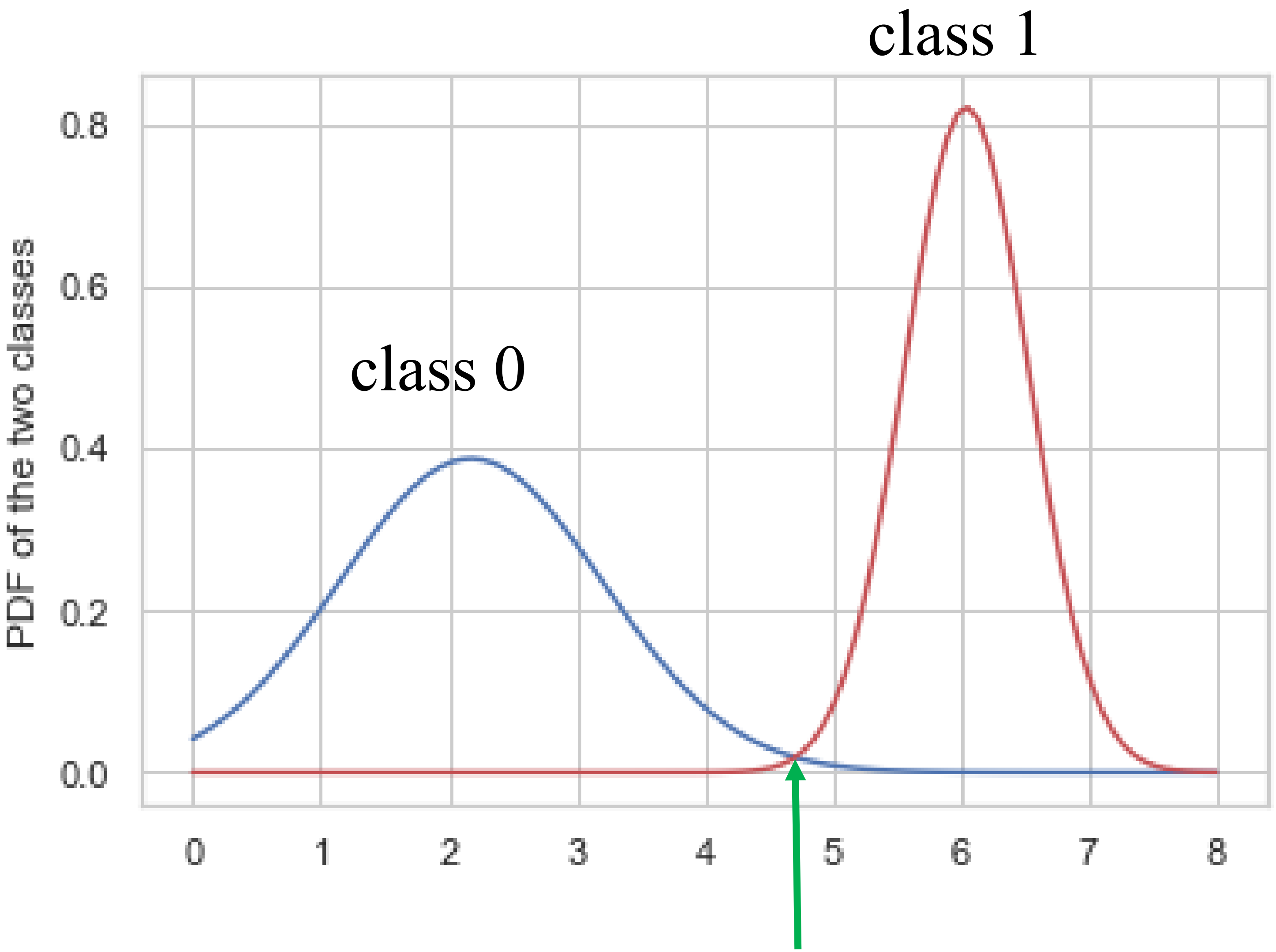
Large error if the two PDFs have large overlap



Error of Non-Bayes decision boundary > Error of Bayes decision boundary



Error of Non-Bayes decision boundary > Error of Bayes decision boundary



Decision Rule:
if x > **t**, then x is from class 1
if x < **t**, then x is from class 0.

when $\pi_0 = \pi_1$=0.5, the decision threshold (**t**) is at the intersection point where

$$\boldsymbol{p(x|y=0)} = \boldsymbol{p(x|y=1)}$$

An application-dependent loss

$$error\ rate = \boldsymbol{\pi_1} \int_{x \to \boldsymbol{C_0}} \boldsymbol{p(x|y=1)}dx + \boldsymbol{\pi_0} \int_{x \to \boldsymbol{C_1}} \boldsymbol{p(x|y=0)}dx$$

$$loss = \boldsymbol{B} \times \boldsymbol{\pi_1} \int_{x \to \boldsymbol{C_0}} \boldsymbol{p(x|y=1)}dx + \boldsymbol{A} \times \boldsymbol{\pi_0} \int_{x \to \boldsymbol{C_1}} \boldsymbol{p(x|y=0)}dx$$

$\boldsymbol{A}$ is the cost of making a wrong decision for data points in class 0

$\boldsymbol{B}$ is the cost of making a wrong decision for data points in class 1

Bayes binary classification: summary

- Assume a PDF $p(x|y)$ for each class (e.g., Gaussian, or GMM)
- Assume a prior distribution $p(y)$ (it could be complex and have many parameters)
- Train the classifier on training data, which is to estimate the parameters of the PDFs and the parameters of the prior distributions by minimizing the NLL loss
- Use the trained classifier to classify a new data point

  $p(y = 0|x)$, $p(y = 1|x)$ and log-likelihood ratio $h(x) = log\,\frac{p(y=0|x)}{p(y=1|x)}$

- If every class PDF is a simple Gaussian, a nice analytical form of the decision boundary can be obtained; and we can calculate the error rate.

Naïve Bayes Classifier

- We have a set of data points $\{x_1, x_2, x_3, \ldots, x_N\}$ and $x_n \in \mathcal{R}^M$
- Each data point has M features

$$x_n = \left[x_{n,1}, x_{n,2}, x_{n,3}, \ldots, x_{n,m}, \ldots, x_{n,M}\right]^T$$

- Drop the index n

$$x = \left[x_{(1)}, x_{(2)}, x_{(3)}, \ldots, x_{(m)}, \ldots, x_{(M)}\right]^T$$

- A Naïve Bayes classifier assumes a PDF such that

$$p(x|y) = p\left(x_{(1)}\middle|y\right)p\left(x_{(2)}\middle|y\right)p\left(x_{(3)}\middle|y\right)\ldots p\left(x_{(M)}\middle|y\right) = \prod_{m=1}^{M} p\left(x_{(m)}\middle|y\right)$$

For each data point x, the feature components are assumed to be independent

It becomes easy to compute the PDF value at each data point x

Naïve Bayes Classifier: 2D Gaussian PDF

- Each data point $x = \left[x_{(1)}, x_{(2)}\right]^T$ is in 2D space
- A Naïve Bayes classifier assumes a PDF such that

$$p(x|y) = p\left(x_{(1)}\middle|y\right)p\left(x_{(2)}\middle|y\right)$$

$$p(x|y) = \left(\frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{x_{(1)}^2}{2\sigma_1^2}}\right)\left(\frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{x_{(2)}^2}{2\sigma_2^2}}\right)$$

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$ which means $x_{(1)}$ *and* $x_{(2)}$ are independent

Bayes_Rule_1D_2D_Gaussian_2Classes.ipynb

What is the difference between
Gaussian Bayes Classifier and GMM ?