CS4487 - Machine Learning

Lecture 2a - Bayes Classifier

Dr. Antoni B. Chan

Dept. of Computer Science, City University of Hong Kong

Outline

1. Bayes Classification and Generative Models
2. Parameter Estimation
3. Bayesian Decision Rule

Classification Examples

- Given an email, predict whether it is spam or not spam.
  - **Email 1:**

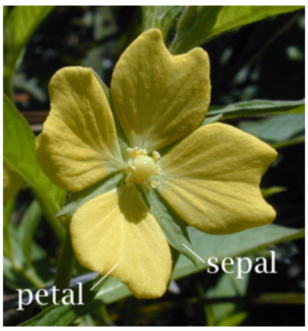
    > There was a guy at the gas station who told me that if I knew Mandarin and Python I could get a job with the FBI.

  - **Email 2:**

    > A home based business opportunity is knocking at your door. Donít be rude and let this chance go by. You can earn a great income and find your financial life transformed. Learn more Here. To Your Success. Work From Home Finder Experts

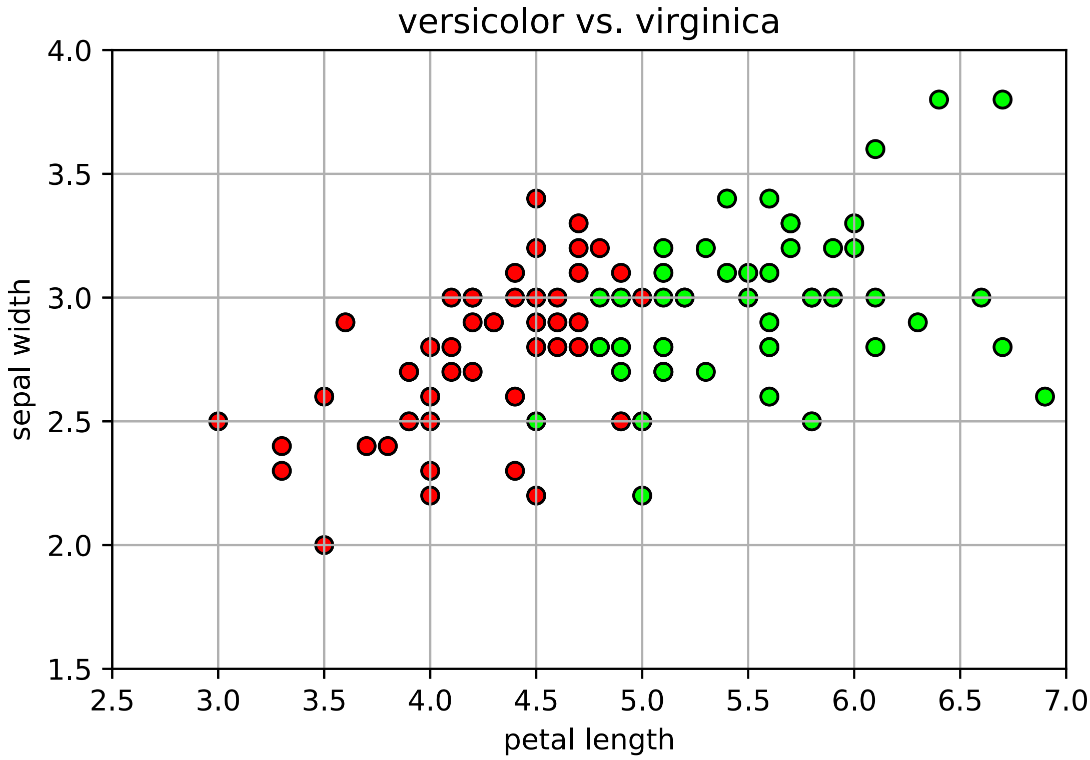
---

- Classification Examples
  - Given the *petal length* and *sepal width*, predict the type of iris flower.

Features	**Versicolor**	**Virginica**
		

```
In [3]: irisfig
```

Out[3]:



General Classification Problem

- Observation $\mathbf{x}$ (i.e., features)
  - typically a real vector, $\mathbf{x} \in \mathbb{R}^d$.
  - **Example**: a 2-dim vector containing the petal length and sepal width.
    - $\mathbf{x} = \begin{bmatrix} \text{petal length} \\ \text{sepal width} \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$
- Class y
  - takes values from a set of possible class labels $\mathcal{Y}$.
  - **Example:** $\mathcal{Y} = \{\text{"versicolor"}, \text{"virginica"}\}$.
    - or equivalently as numbers, $\mathcal{Y} = \{1, 2\}$.
- **Goal**: given an observed features $\mathbf{x}$, predict its class y.

Probabilistic model

- One type of classifier is to model the data.
- Model *how* the data is generated using probability distributions.
  - called a **generative model**.
- Generative model
  - 1) The world has objects of various classes.
  - 2) The observer measures features/observations from the objects.
  - 3) Each class of objects has a particular distribution of features.

Class model

- possible classes are $\mathcal{Y}$
  - for example, $\mathcal{Y} = \{\text{"versicolor"}, \text{"virginica"}\}$.
    - or more generally, $\mathcal{Y} = \{1, 2\}$.
- in the world, the frequency that class y occurs is given by the probability distribution $p(y)$.
  - $p(y)$ is called the **prior distribution.**

- **Example:**
  - $p(y = 1) = 0.4$
  - $p(y = 2) = 0.6$
  - "In the world of iris flowers, there are 40% that are Class 1 (versicolor) and 60% that are Class 2 (virginica)"

Learn from our data

- $p(y = 1) = \frac{\text{number of examples of Class 1}}{\text{total number of examples}}$
- analogous for Class 2

In [4]:
```
N1 = count_nonzero(y==1)  # number of Class 1 examples
N2 = count_nonzero(y==2)  # number of Class 2 examples
N  = len(y)               # total
py = [double(N1)/N, double(N2)/N] # note: avoids integer division!
print(py)
```

```
[0.5, 0.5]
```

Observation model

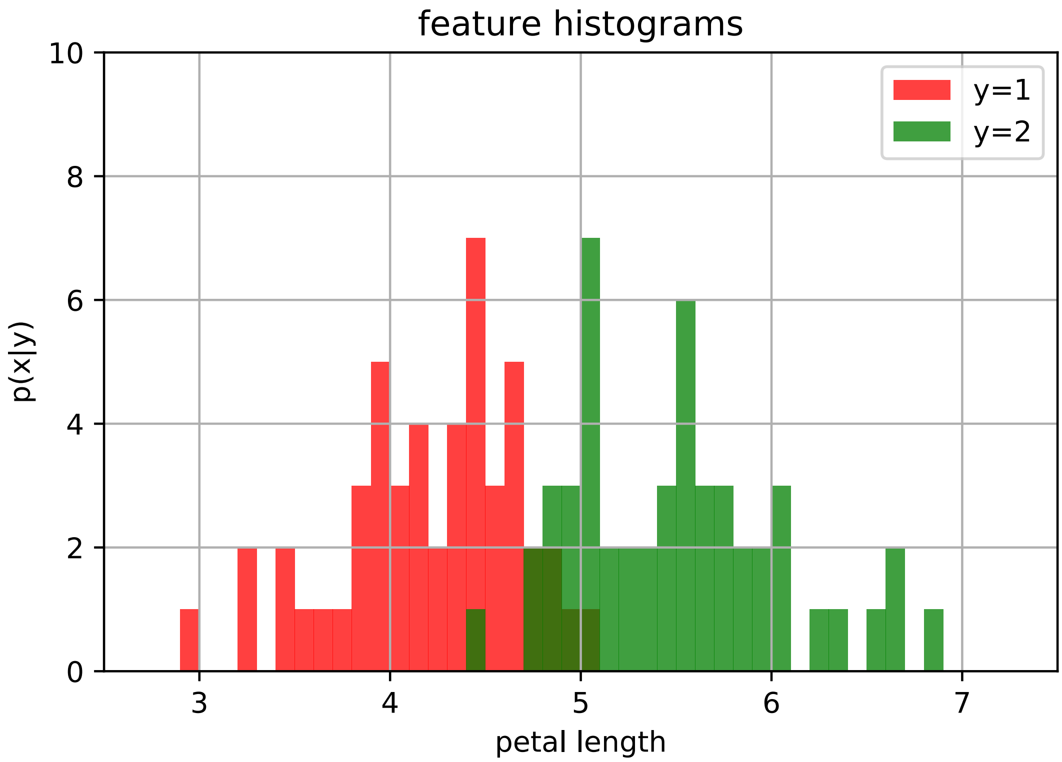
- we measure/observe feature vector $\mathbf{x}$
  - the value of the features *depend* on the class.
- the observation is drawn according to the distribution $p(\mathbf{x}|y)$.
  - $p(x|y)$ is called the **class conditional distribution**
    - "probability of observing a particular feature vector $\mathbf{x}$ given the object is class y"
    - can "smooth out the samples" or "fill-in" values between samples.

Learn from the data

- histograms for feature "petal length" for each class

In [6]:
```
ccdhist
```

Out[6]:



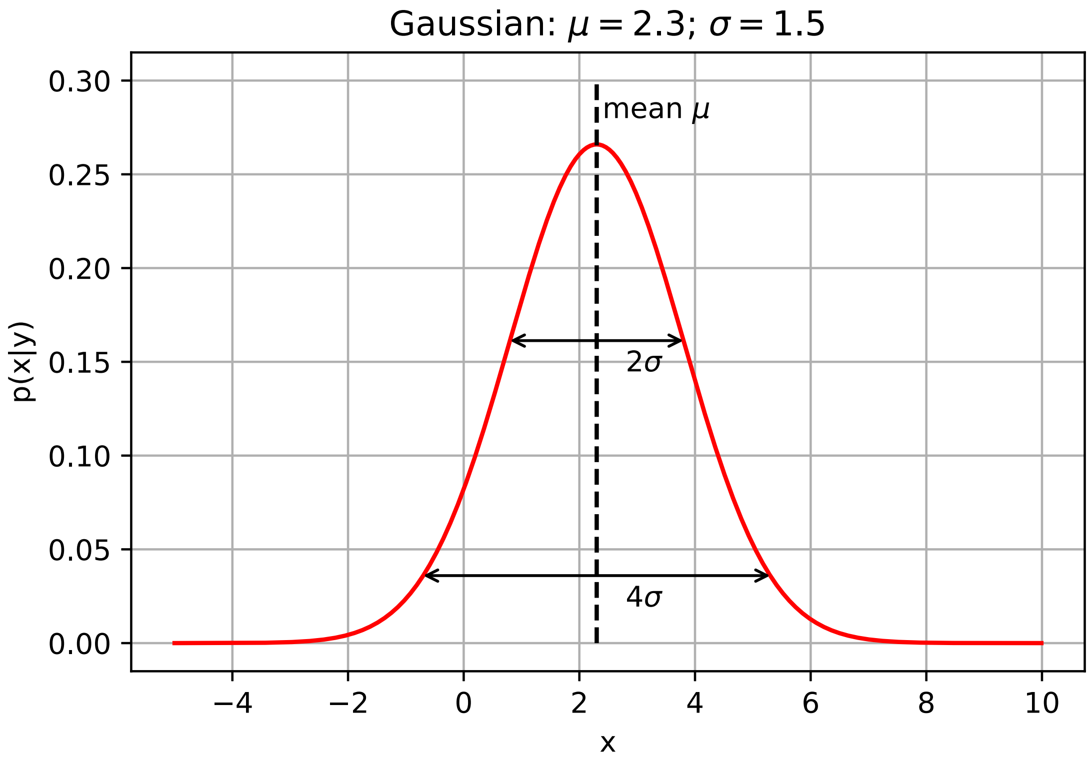
- **Problem:** looks a little bit noisy.
- **Solution:** assume a probability model for the class conditional $p(x|y)$

Gaussian distribution (normal distribution)

- Each class is modeled as a separate Gaussian distribution of the feature value
  - $p(x|y=c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} e^{-\frac{1}{2\sigma_c^2}(x-\mu_c)^2}$
  - Each class has its own mean and variance parameters (μ_c, σ_c^2).

In [8]:
```
gfig
```

Out[8]:



Learn the parameters from data.

- Maximum likelihood estimation (MLE)
  - set the parameters (μ, σ^2) to maximize the likelihood (probability) of the samples for that class.
  - Let $\{\mathbf{x}_i, y_i\}_{i=1}^N$ be the data for one class:

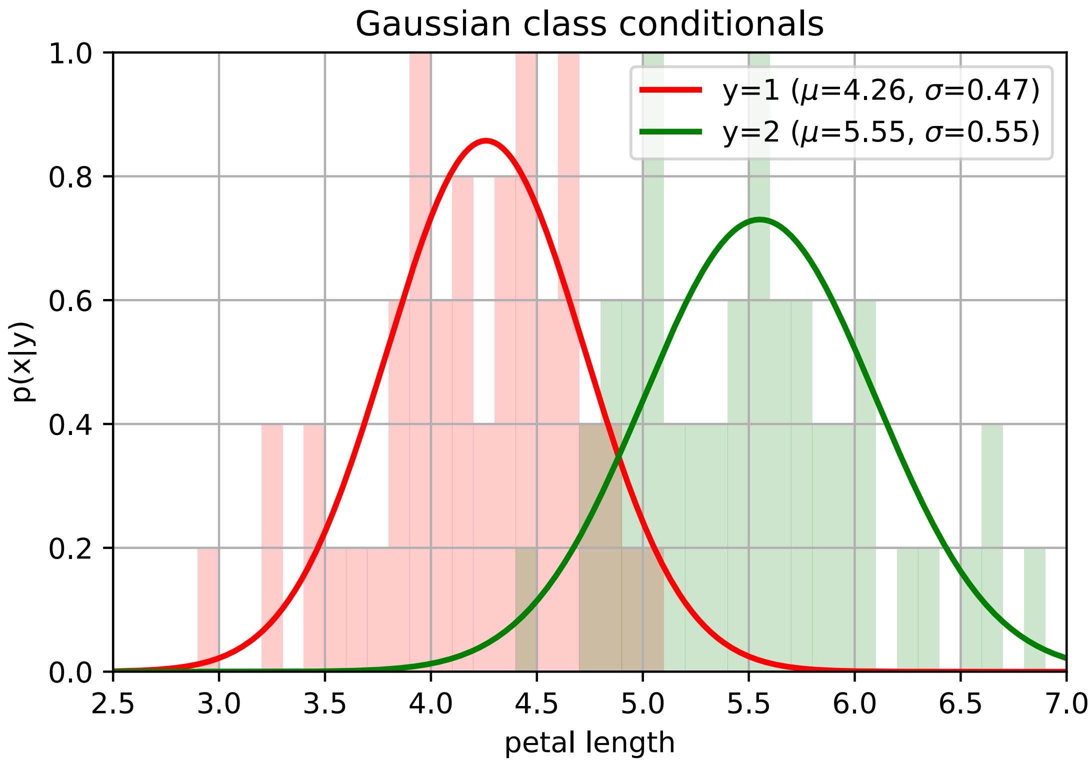
$$(\hat{\mu}, \hat{\sigma^2}) = \operatorname*{argmax}_{\mu, \sigma^2} \sum_{i=1}^{N} \log p(\mathbf{x}_i | y_i)$$

- Solution:
  - sample mean: $\hat{\mu} = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i$
  - sample variance: $\hat{\sigma^2} = \frac{1}{N} \sum_{i=1}^N (\mathbf{x}_i - \hat{\mu})^2$

In [11]:

```
gcd
```

Out[11]:



Bayesian Decision Rule

- The Bayesian decision rule (BDR) makes the optimal decisions on problems involving probability (uncertainty).
  - minimizes the *probability of making a prediction error*.
- **Bayes Classifier**
  - Given observation x, pick the class c with the *largest posterior probability*, $p(y = c|x)$.
  - **Example:**
    - if $p(y = 1|x) > p(y = 2|x)$, then choose Class 1
    - if $p(y = 1|x) < p(y = 2|x)$, then choose Class 2

- Problem: we don't have $p(y|x)$!
  - we only have $p(y)$ and $p(x|y)$.

Bayes' Rule

- The posterior probability can be calculated using Bayes' rule:

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

  - The denominator is the probability of x:
    - $p(x) = \sum_{y \in \mathcal{Y}} p(x|y)p(y)$
  - The denominator makes $p(y|x)$ sum to 1.

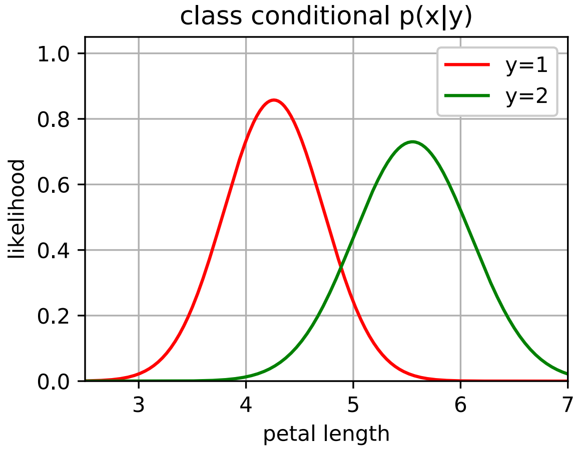
- Bayes' rule:

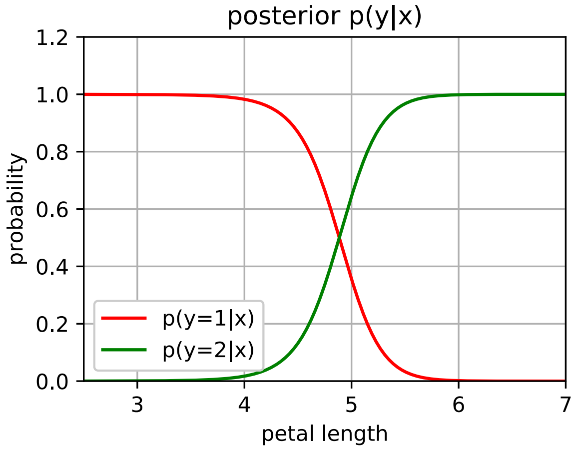
$$p(y|x) = \frac{p(x|y)p(y)}{p(x|y=1)p(y=1) + p(x|y=2)p(y=2)}$$

- **Example**:

In [13]:
```
iris1dpost
```

Out[13]:

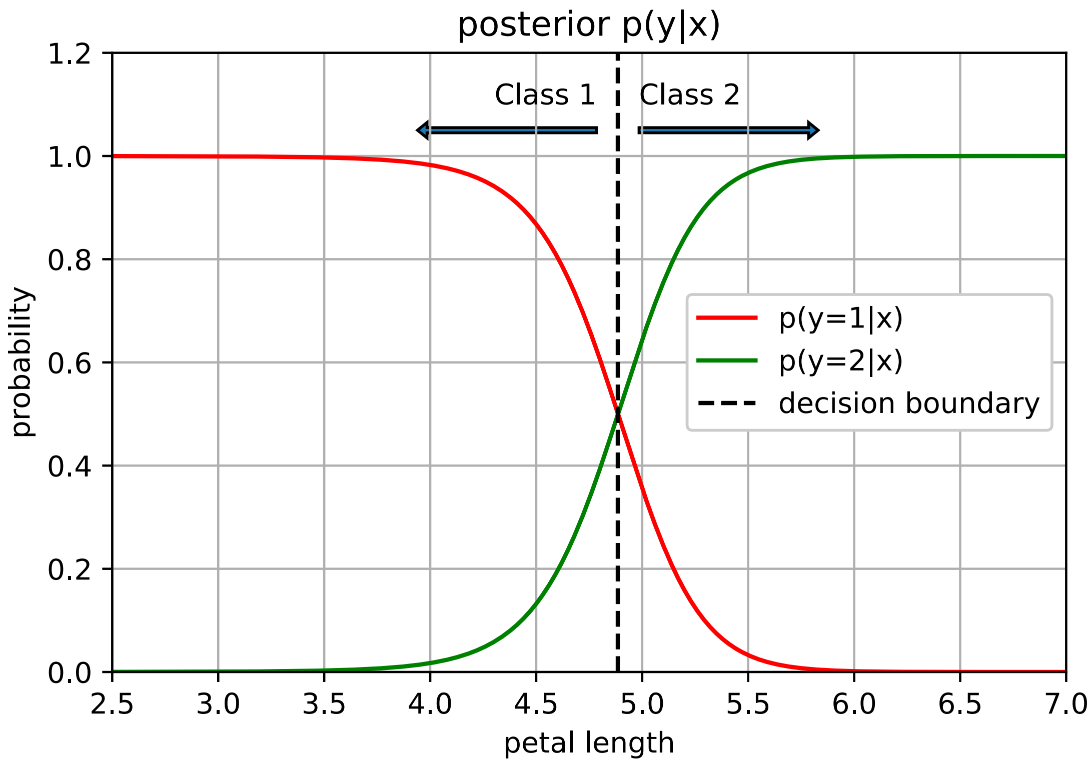




- The *decision boundary* is where the two posterior probabilites are equal
  - $p(y=1|x) = p(y=2|x)$

In [15]:
```
iris1dpost2
```

Out[15]:



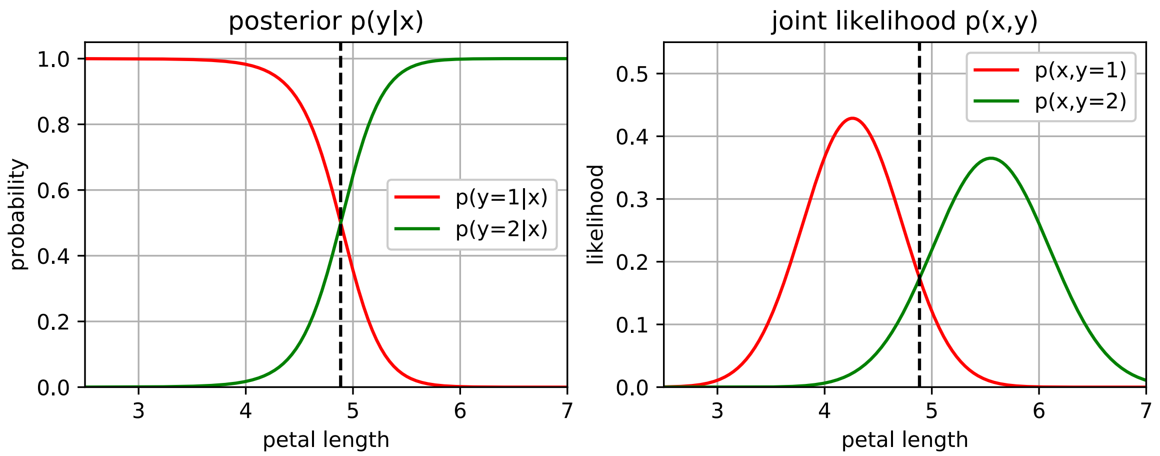
Bayes rule revisited

- Bayes' rule: $p(y|x) = \frac{p(x|y)p(y)}{p(x)}$
- Note that the denominator is the same for each class y.
  - hence, we can compare just the numerators $p(x|y)p(y)$.
  - This also called the *joint likelihood* of the observation and class
    - $p(x, y) = p(x|y)p(y)$

- **Example:**
  - BDR using joint likelihoods:
    - if $p(x|y = 1)p(y = 1) > p(x|y = 2)p(y = 2)$, then choose Class 1
    - otherwise, choose Class 2

In [17]:
```
iris1djoint
```

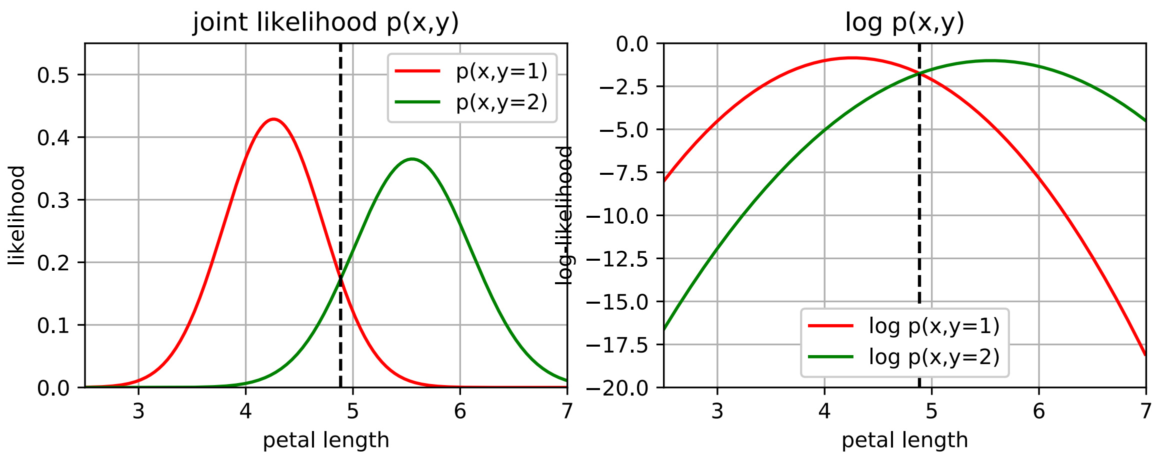
Out[17]:



- Can also apply a monotonic increasing function (like $\log$) and do the comparison.
  - Using log likelihoods:
    - $\log p(x|y = 1) + \log p(y = 1) > \log p(x|y = 2) + \log p(y = 2)$
  - This is more numerically stable when the likelihoods are small.

In [19]:
```
iris1dLL
```

Out[19]:



Bayes Classifier Summary

- **Training:**
  1. Collect training data from each class.
  2. For each class c, estimate the class conditional densities $p(x|y=c)$:
     A. select a form of the distribution (e.g. Gaussian).
     B. estimate its parameters with MLE.
  3. Estimate the class priors $p(y)$ using MLE.
- **Classification:**
  1. Given a new sample x^*, calculate the likelihood $p(x^*|y=c)$ for each class c.
  2. Pick the class c with largest posterior probability $p(y=c|x)$.
     - (equivalently, use $p(x|y=c)p(y=c)$ or $\log p(x|y=c) + \log p(y=c)$)