Bayesian Decision Theory

Instructor: Tom Ko

Things that assume you know

- Basic probability theory

Bayes' theorem

Bayes' theorem is stated mathematically as the following equation:

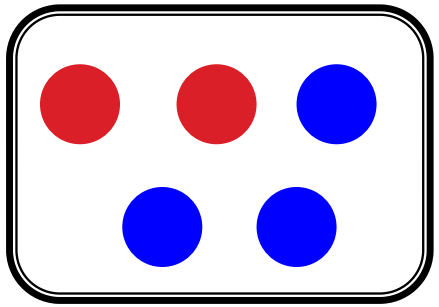
$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

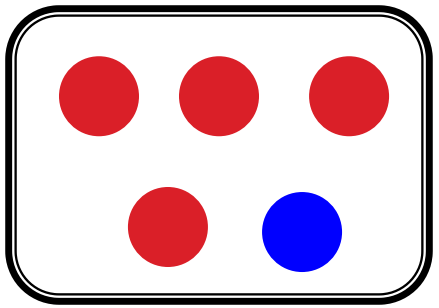
where A and B are events and $P(B) \neq 0$.

- P(A|B) is a conditional probability: the likelihood of event A occurring given that B is true
- P(A) and P(B) are the probabilities of observing A and B respectively. They are known as the marginal probability.

Basic Concepts of Probability

Bag 1



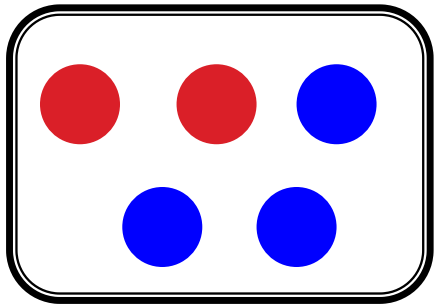


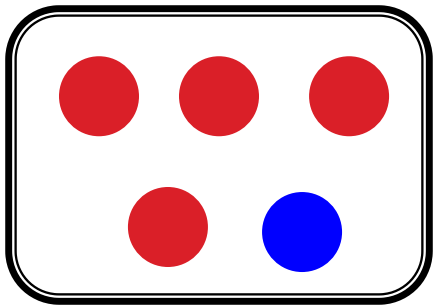
Bag 2

- P(red|bag1) = 0.4
- P(red|bag2) = 0.8
- The problem: given that a red ball is drawn, can you guess which bag the ball is drawn from?
- P(bag1|red) or P(bag2|red) has a larger value?

Consider the prior

Bag 1



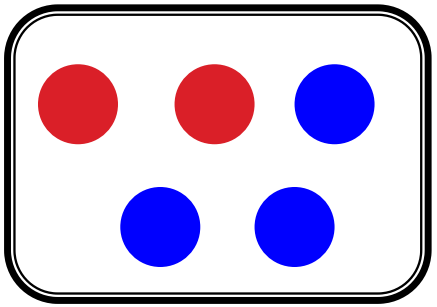


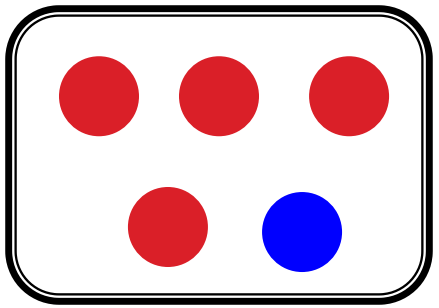
Bag 2

- Given P(bag1) = 0.7 and P(bag2) = 0.3
- P(red|bag1) = 0.4
- P(red|bag2) = 0.8
- P(bag1|red) or P(bag2|red) has a larger value?

Applying Bayes' theorem

Bag 1





Bag 2

- Given P(bag1) = 0.7 and P(bag2) = 0.3
- P(red|bag1) = 0.4
- P(red|bag2) = 0.8

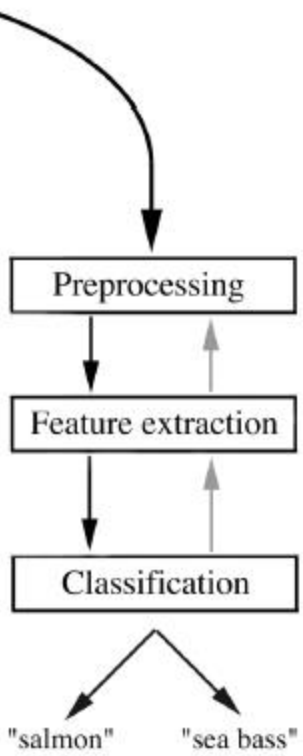
$$P(bag1|red) = \frac{P(red|bag1)\,P(bag1)}{P(red)} = \frac{0.4 * 0.7}{P(red)} = \frac{0.28}{P(red)}$$

$$P(bag2|red) = \frac{P(red|bag2)\,P(bag2)}{P(red)} = \frac{0.8 * 0.3}{P(red)} = \frac{0.24}{P(red)}$$

The problem

- A fish packing company wants to automate the process of sorting incoming fish.
- Two types of fish: Sea bass and Salmon
- How do you design a classifier for it?
- What you can get is the weight and length of each fish.

Making the process automatic



Making a naive classifier

- Using the length to classify

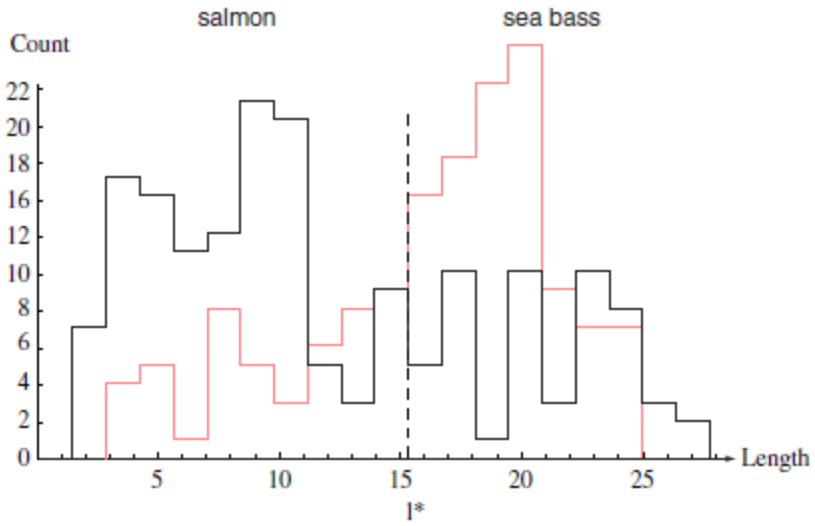


Figure 1.2: Histograms for the length feature for the two categories. No single threshold value l^* (decision boundary) will serve to unambiguously discriminate between the two categories; using length alone, we will have some errors. The value l^* marked will lead to the smallest number of errors, on average.

Making a naive classifier

- Using the weight to classify

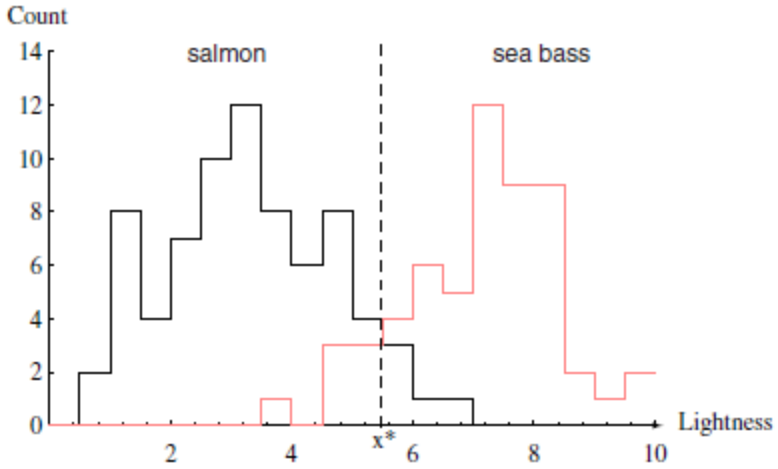


Figure 1.3: Histograms for the lightness feature for the two categories. No single threshold value x^* (decision boundary) will serve to unambiguously discriminate between the two categories; using lightness alone, we will have some errors. The value x^* marked will lead to the smallest number of errors, on average.

Making a naive classifier

- Using both features

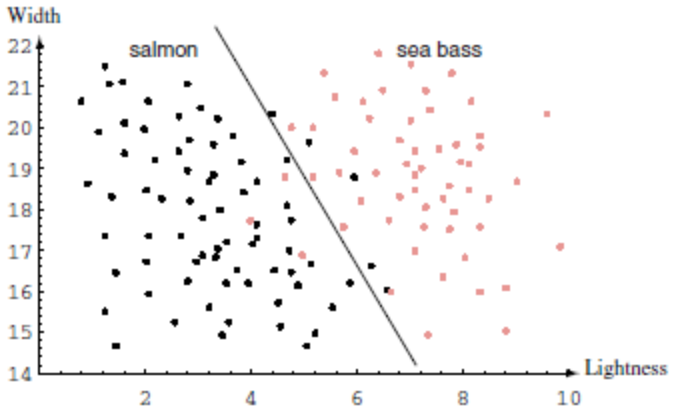


Figure 1.4: The two features of lightness and width for sea bass and salmon. The dark line might serve as a decision boundary of our classifier. Overall classification error on the data shown is lower than if we use only one feature as in Fig. 1.3, but there will still be some errors.

Define the problem

- The fish problem is a two class classification problem
- let ω denote the class, with $\omega = \omega_1$ for sea bass and $\omega = \omega_2$ for salmon.

Make decision with prior

- Rule: Decide $\omega 1$ if $P(\omega_1) > P(\omega_2)$; otherwise decide ω_2
- Make sense if to judge only one fish
- Strange if we have to judge many fish and if $P(\omega_1)$ and $P(\omega_2)$ are close to each other.

Bayes decision rules

- Rule : Decide ω_1 if $P(\omega_1|x) > P(\omega_2|x)$; otherwise decide ω_2
- This emphasizes the role of the posterior probabilities.
- Please note that $P(\omega_1|x) + P(\omega_2|x) = 1$.

Bayes formula

- $p(\omega_j, x) = P(\omega_j / x)p(x) = p(x/\omega_j)P(\omega_j)$
- Rearranging these leads us to the answer to our question, which is called *Bayes' formula*:

$$\boxed{P(\omega_j|x) = \frac{p(x|\omega_j)P(\omega_j)}{p(x)},}$$

$$posterior = \frac{likelihood \times prior}{evidence}$$

- *Here,* the evidence factor, $p(x)$, is not important to the decision making

$$p(x) = \sum_{j=1}^{2} p(x|\omega_j)P(\omega_j)$$

Class-conditional probability density function

- We call $p(x/\omega_j)$ the *likelihood* of ω_j with respect to x
- It is a probability density function (pdf) of x
- What if we know the feature of the fish?
- We can use the *class-conditional probability density* function $p(x/\omega_1)$ and $p(x/\omega_2)$ to improve our classifier.
- Here we consider x to be a continuous random variable

Class-conditional probability density function

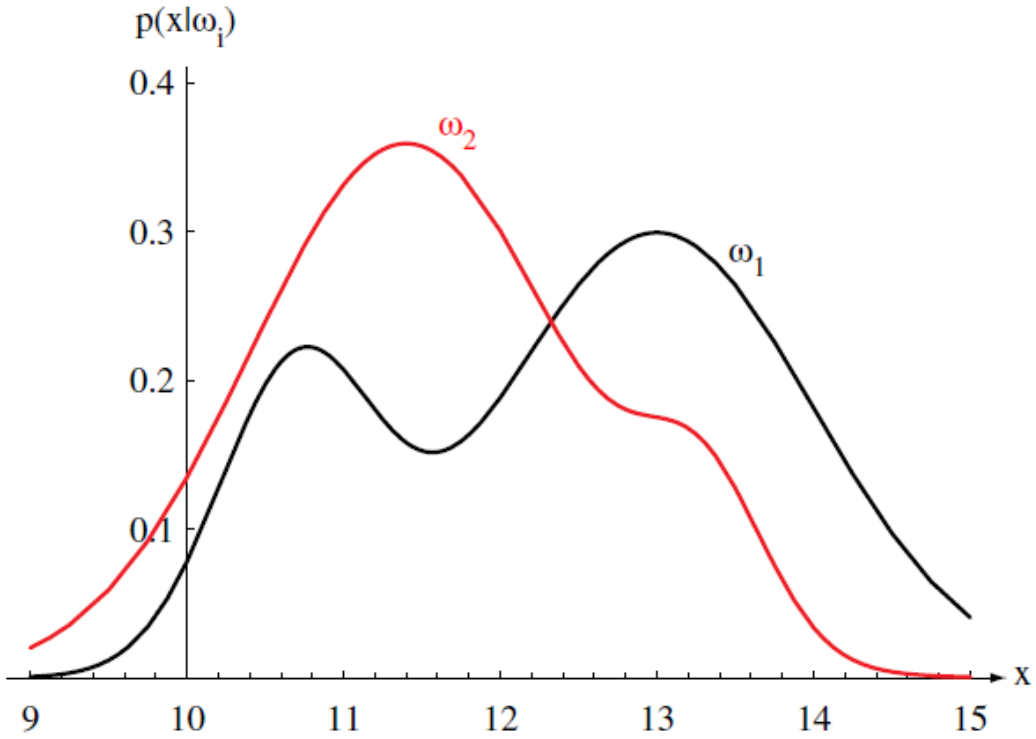


Figure 2.1: Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x given the pattern is in category ω_i. If x represents the length of a fish, the two curves might describe the difference in length of populations of two types of fish. Density functions are normalized, and thus the area under each curve is 1.0.

Prior

- If the catch produced as much sea bass as salmon, we would say that the next fish is equally likely to be sea bass or salmon.
- You may want to make a guess before knowing any information of the next fish.
- There is *a priori probability* $P(\omega_1)$ that the next fish is sea bass, and a prior probability $P(\omega_2)$ that it is salmon.

Posterior

- $\underset{j}{\operatorname{argmax}} P(\omega_j|x)$ where $P(\omega_j/x)$ sum up to one for all j
- Bayes' formula shows that by observing the value of x we can convert the prior probability $P(\omega_j)$ to the *posterior* probability $P(\omega_j/x)$

The posterior curve

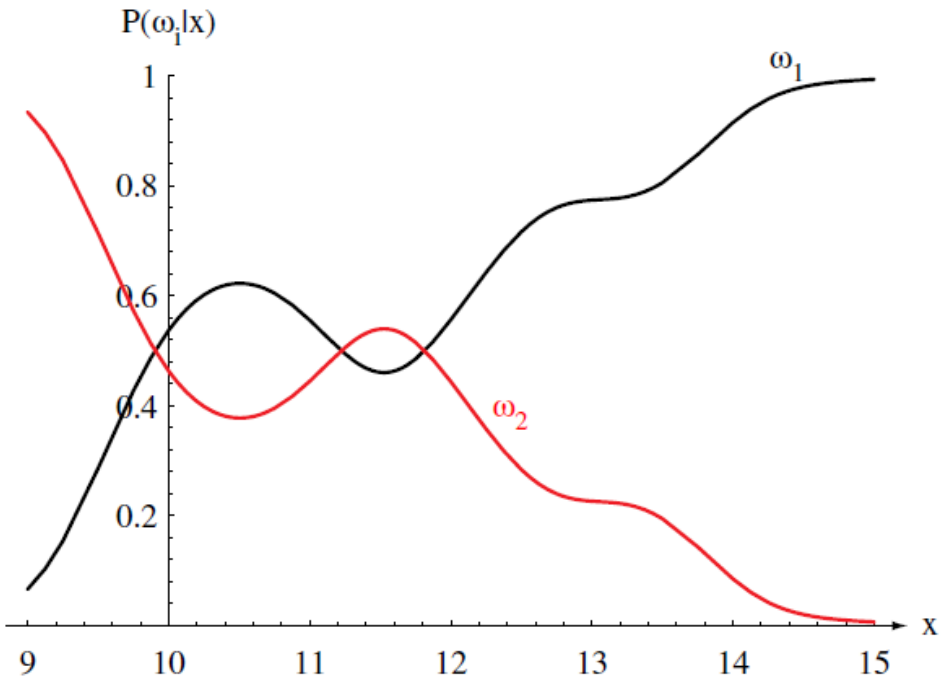


Figure 2.2: Posterior probabilities for the particular priors $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$ for the class-conditional probability densities shown in Fig. 2.1. Thus in this case, given that a pattern is measured to have feature value $x = 14$, the probability it is in category ω_2 is roughly 0.08, and that it is in ω_1 is 0.92. At every x, the posteriors sum to 1.0.

To learn the classifier

- The structure of a Bayes classifier is determined by the conditional densities ($p(x/\omega_1)$, $p(x/\omega_2)$) as well as by the prior probabilities ($p(\omega_1)$, $p(\omega_2)$).
- How to estimate all of them?
- The way to estimate is the learning process.
- Where to learn from?
  - To learn from the data.

Statistical learning

- We have no where to learn from except from the data itself.
- Data is infinite, we can only learn from the samples we got.
- Can we catch all the sea bass and salmon of world to estimate the figures?
- The more samples you got, the more accurate is the estimate.

Learning the prior probabilities

- Just count
- Let say we catch 100 fishes. Count how many of them are sea bass and how many are salmon.
- If 40 fishes are sea bass, $P(\omega_1) = 40/100 = 0.4$, then $P(\omega_2) = 0.6$
- We use these figures to represent the underlying global probability of sea bass and salmon of the whole world.

Learning the conditional PDF

- What exactly is $p(x/\omega_1)$?
  - It is a function of x
- x is a continuous random variable
- Every random variable has a probability distribution
- In our case, x is the length of the fish, we believe that the generation of x is governed by the underlying PDF
- If we can have every fish samples in the world, we can know the underlying PDF by recording every fish length.
- However, we can only estimate the underlying PDF from limited samples.

Learning the conditional PDF

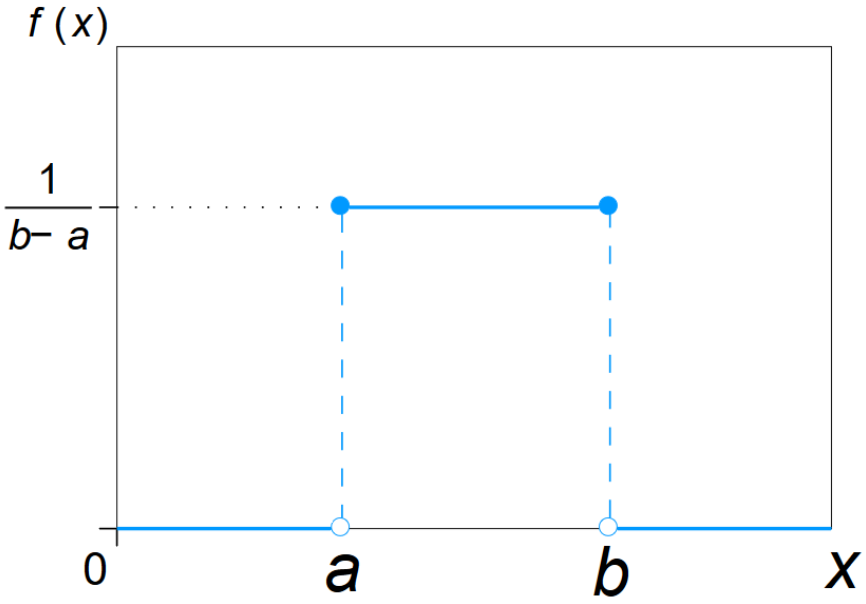
- We need to give $p(x/\omega_1)$ a functional form.
- There are various density functions
  - Uniform distribution
  - Gaussian distribution
- Give an assumption to the underlying PDF according to the situation.

Bus waiting time

- You show up at a bus stop to wait for a bus that comes by once per hour.
- You do not know what time the bus came by last.
- If the last bus just left, you need to wait long.
- If the last bus have left for almost an hour, very soon the next bus should come.

Uniform distribution

- The arrival time of the next bus is a continuous uniform distribution [0,1] measured in hours.



Human height

- What is the distribution of human height likely to be?
- Most of the people in a specific population are of average height.
- A very small number of people are either extremely tall or extremely short.

The Normal Density

- Also known as Gaussian distribution
- The variance affects the flatness of the curve

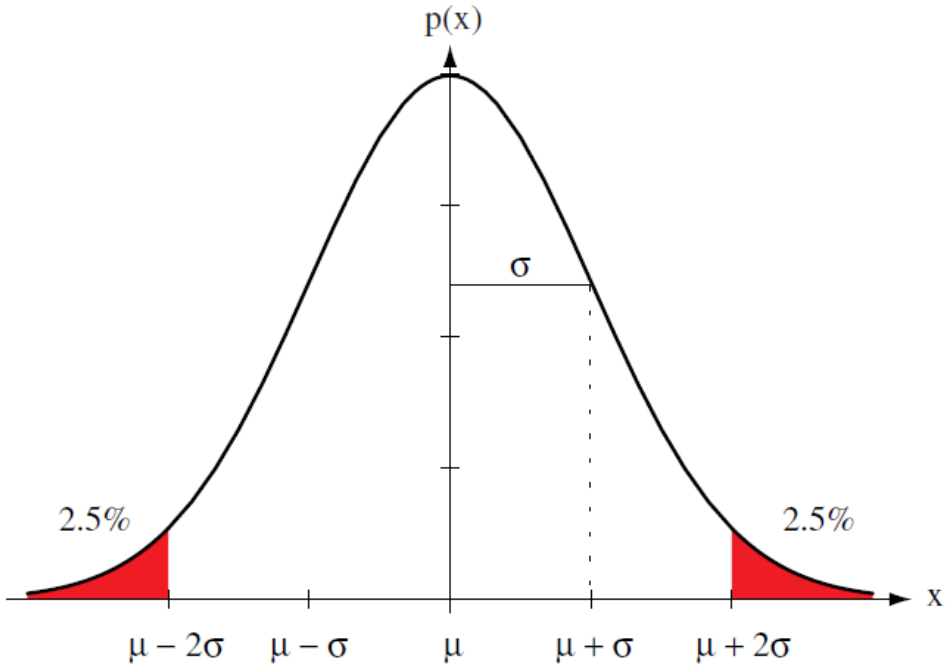


Figure 2.7: A univariate normal distribution has roughly 95% of its area in the range $|x - \mu| \leq 2\sigma$, as shown. The peak of the distribution has value $p(\mu) = 1/\sqrt{2\pi}\sigma$.

Univariate Gaussian

- The basic form

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

- The univariate normal density is completely specified by two parameters: its mean μ and variance σ^2
- For simplicity, write $p(x) \sim N(\mu, \sigma^2)$

Using the classifier

- Comparing P(ω_1|x) and P(ω_2 |x)

$$P(\omega_j|x) = \frac{p(x|\omega_j)P(\omega_j)}{p(x)},$$

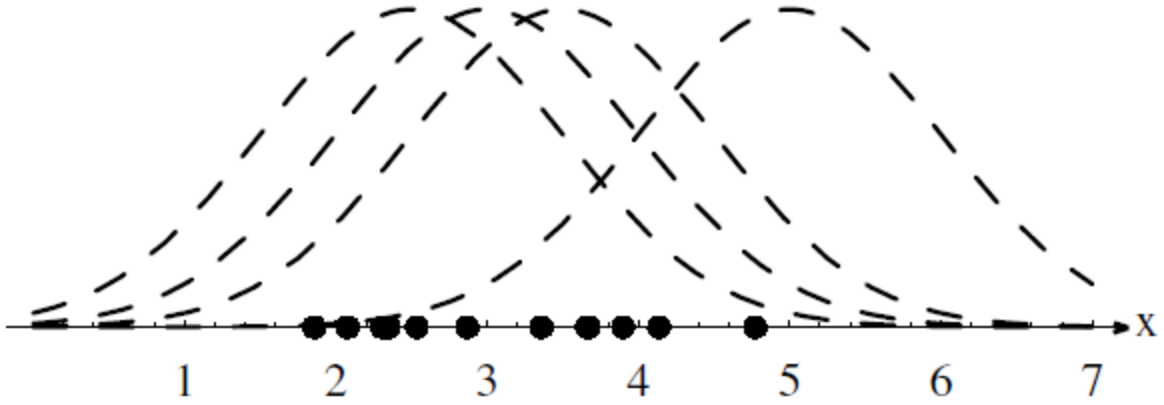
- We are actually comparing P(x|ω_1)P(ω_1) and P(x|ω_2)P(ω_2)
- For class ω1, the model parameters are μ_1 and σ_1, thus

$$p(x|\omega_1) = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left[-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2\right]$$

- Similar for class ω_2

Estimating the parameters

- Now the problem is, what are the exact values of μ and σ^2 should be?
- We should estimate them from the data



Estimating the parameters

- Formally define the way of estimation
  - Maximum likelihood estimation
  - Bayesian estimation

Maximum likelihood(ML) estimation

- Viewing the parameters as quantities whose values are fixed but unknown.
- The best estimate of the values is defined to be those maximizing the probability of obtaining the samples actually observed.

ML – The general principle

- Suppose a collection of samples D, $\mathbf{x}_1$, …, $\mathbf{x}_n$, *is* drawn according to $p(x/\omega_1)$ (PDF of class 1).
- Such samples are *i.i.d.* — independent identically distributed random variables.
- Assume that $p(\mathbf{x}/\omega_1)$ has a known parametric form, say, a Gaussian distribution $N(\mu, \sigma^2)$
- Our goal is to use the information provided by the training samples to obtain good estimates for the unknown parameter vectors $\boldsymbol{\theta}_1$.
- In the Gaussian case, $\boldsymbol{\theta}_1$ contains μ and σ^2

ML – The general principle

- Suppose the samples are drawn independently, we have

$$p(\mathcal{D}|\theta) = \prod_{k=1}^{n} p(\mathbf{x}_k|\theta)$$

- $p(D/\theta)$ is called the *likelihood* of θ with respect to the set of samples.
- The *maximum likelihood estimate* of θ is, by definition, the value that maximizes $p(D/\theta)$.

$$\hat{\theta} = \arg\max_{\theta} p(\mathcal{D}|\theta)$$

ML – The general formulas

$$\hat{\theta} = \arg\max_{\theta} p(\mathcal{D}|\theta)$$

$$\hat{\theta} = \arg\max_{\theta} l(\theta) \longrightarrow l(\theta) \equiv \ln p(\mathcal{D}|\theta) \longrightarrow p(\mathcal{D}|\theta) = \prod_{k=1}^{n} p(\mathbf{x}_k|\theta)$$

$$\nabla_{\theta} l = 0 \qquad l(\theta) = \sum_{k=1}^{n} \ln p(\mathbf{x}_k|\theta)$$

$$\nabla_{\theta} l = \sum_{k=1}^{n} \nabla_{\theta} \ln p(\mathbf{x}_k|\theta)$$

The Gaussian Case: Unknown μ

$$\nabla_{\boldsymbol{\theta}} l = \sum_{k=1}^{n} \nabla_{\boldsymbol{\theta}} \ln p(\mathbf{x}_k|\boldsymbol{\theta})$$

$$\nabla_{\mu} l = \sum_{k=1}^{n} \nabla_{\mu} \ln p(\mathbf{x}_k|\mu) \longrightarrow p(x|\mu) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

$$\ln p(x|\mu) = -\frac{1}{2}\ln[2\pi\sigma^2] - \frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2$$

$$\nabla_{\mu} \ln p(x|\mu) = \frac{x-\mu}{\sigma^2}$$

$$\nabla_{\mu} l = 0$$

$$\sum_{k=1}^{n} \frac{\mathbf{x}_k - \hat{\mu}}{\sigma^2} = 0 \quad \Rightarrow \quad \hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k$$

The Gaussian Case: Unknown μ and σ

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \hat{\mu})^2$$

- Please solve the variance case on your own.

Maximum a posteriori (MAP) estimation

- China wants to estimate the distribution of the height of all citizens in the country. We believed that it is a normal distribution $p(x|\mu) \sim N(\mu, \sigma^2)$
- By drawing samples of human height in the country, we could estimate μ by ML.
- However, it was told the distribution of average human height of other countries. $p(\mu) \sim N(\mu_0, \sigma_0^2)$

  ,where μ_0 is the average of these average heights

- Thus, we can make use of this prior distribution of the parameters in estimation.

Maximum a posteriori (MAP) estimation

- If you know some prior knowledge about θ, say P(θ), we can use it to help the estimation.

ML: $$\hat{\theta} = \arg\max_{\theta} l(\theta)$$

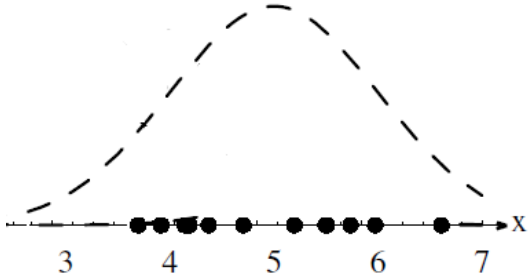
MAP: $$\hat{\theta} = \arg\max_{\theta} l(\theta)\, p(\theta)$$

Bayesian estimation

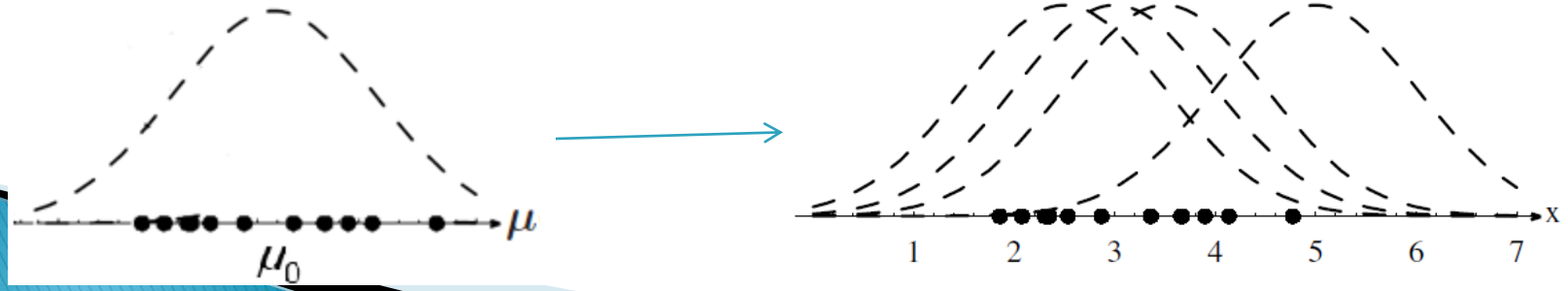
- In Bayesian learning, we view the parameters as random variables having some known a priori distribution.
- It wants to consider every possibility of θ

Bayesian vs. ML (Gaussian case)

- In ML's view, the parameter vector **θ** is fixed.
  - In the Gaussian case, μ governs the distribution of x



- In Bayesian's view , the parameter vector **θ** is a random variable.
  - In the Gaussian case, μ_0 governs the distribution of μ and μ governs the distribution of x



In the view of data generation

- Non-Bayesian: assuming that the parameters are fixed values.
- E.g., the observed data are generated by a Gaussian with a particular μ

- Bayesian: assuming that the parameters are random variables.
- E.g., the observed data are generated by a Gaussian with a μ, where μ is randomly generated.
- You can imagine the data are generated with a lot of Gaussian with different μ.

In the view of classification

- For classification purpose, our goal is to compute

$$P(\omega_i|\mathbf{x})$$

- The information resides in the training samples:

$$P(\omega_i|\mathbf{x},\mathcal{D}) = \frac{p(\mathbf{x}|\omega_i,\mathcal{D})P(\omega_i|\mathcal{D})}{\sum_{j=1}^{c} p(\mathbf{x}|\omega_j,\mathcal{D})P(\omega_j|\mathcal{D})}$$

- The formula refines to

$$P(\omega_i|\mathbf{x},\mathcal{D}) = \frac{p(\mathbf{x}|\omega_i,\mathcal{D}_i)P(\omega_i)}{\sum_{j=1}^{c} p(\mathbf{x}|\omega_j,\mathcal{D}_j)P(\omega_j)}$$

Bayesian estimation

- The central problem of Bayesian learning is to determine

$$p(\mathbf{x}|\omega_i, \mathcal{D}_i)$$

- Since we can treat each class independently, the class notation can be simplified. The central problems becomes: use a set of samples to determine

$$p(\mathbf{x}|\mathcal{D})$$

$$p(\mathbf{x}|\mathcal{D}) = \int p(\mathbf{x}, \theta|\mathcal{D})\ d\theta$$

$$p(\mathbf{x}|\mathcal{D}) = \int p(\mathbf{x}|\theta)p(\theta|\mathcal{D})\ d\theta$$

The Gaussian Case: Unknown μ

$$p(\mu|\mathcal{D}) = \frac{p(\mathcal{D}|\mu)p(\mu)}{\int p(\mathcal{D}|\mu)p(\mu)\, d\mu}$$

$$= \alpha \prod_{k=1}^{n} p(x_k|\mu)p(\mu)$$

$$p(x|\mu) \sim N(\mu, \sigma^2)$$

$$p(\mu) \sim N(\mu_0, \sigma_0^2)$$

$$p(\mu|\mathcal{D}) = \alpha \prod_{k=1}^{n} \overbrace{\frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x_k-\mu}{\sigma}\right)^2\right]}^{p(x_k|\mu)} \overbrace{\frac{1}{\sqrt{2\pi}\sigma_0} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right]}^{p(\mu)}$$

$$= \alpha' \exp\left[-\frac{1}{2}\left(\sum_{k=1}^{n}\left(\frac{\mu-x_k}{\sigma}\right)^2 + \left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right)\right]$$

$$= \alpha'' \exp\left[-\frac{1}{2}\left[\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2 - 2\left(\frac{1}{\sigma^2}\sum_{k=1}^{n} x_k + \frac{\mu_0}{\sigma_0^2}\right)\mu\right]\right]$$

The Gaussian Case: Unknown μ

$$p(\mu|\mathcal{D}) = \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_n}{\sigma_n}\right)^2\right]$$

,where $$\mu_n = \left(\frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2}\right)\bar{x}_n + \frac{\sigma^2}{n\sigma_0^2+\sigma^2}\mu_0$$

$$\sigma_n^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2+\sigma^2}$$

The Gaussian Case: Unknown μ

- Now we got the posterior density for the mean $p(\mu/D)$, we can go to the "class-conditional" density which we need in the classifier.

$$
\begin{aligned}
p(x|\omega_j, \mathcal{D}_j) = p(x|\mathcal{D}) &= \int p(x|\mu)p(\mu|\mathcal{D})\, d\mu \\
&= \int \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_n}{\sigma_n}\right)^2\right] d\mu \\
&= \frac{1}{2\pi\sigma\sigma_n} \exp\left[-\frac{1}{2}\frac{(x-\mu_n)^2}{\sigma^2+\sigma_n^2}\right] f(\sigma, \sigma_n),
\end{aligned}
$$

,where $f(\sigma, \sigma_n) = \int \exp\left[-\frac{1}{2}\frac{\sigma^2+\sigma_n^2}{\sigma^2\sigma_n^2}\left(\mu - \frac{\sigma_n^2 x + \sigma^2\mu_n}{\sigma^2+\sigma_n^2}\right)^2\right] d\mu$

The Gaussian Case: Unknown μ

$$p(x|\omega_j, \mathcal{D}_j) = p(x|\mathcal{D}) \sim N(\mu_n, \sigma^2 + \sigma_n^2)$$

- In effect, the conditional mean μ_n is treated as if it were the true mean, and the known variance is increased to account for the additional uncertainty in x resulting from our lack of exact knowledge of the mean μ
- After all these painful derivation, we can go back to compute the posterior probability $P(\omega_i|\mathbf{x}, \mathcal{D})$ for classification.

Bayesian estimation summarization

- The basic assumptions are summarized as follows:
  - The form of the density $p(\mathbf{x}/\boldsymbol{\theta})$ is assumed to be known, but the value of the parameter vector $\boldsymbol{\theta}$ is not known exactly.
  - Our initial knowledge about $\boldsymbol{\theta}$ is assumed to be contained in a known a priori density $p(\boldsymbol{\theta})$.
  - The rest of our knowledge about $\boldsymbol{\theta}$ is contained in a set D of n samples $\mathbf{x}_1$, ..., $\mathbf{x}_n$ drawn independently according to the unknown probability density $p(\mathbf{x})$.

Generalization

- generalize the classifier in several ways:
  - by allowing the use of more than one feature
  - by allowing more than two classes

Using more features

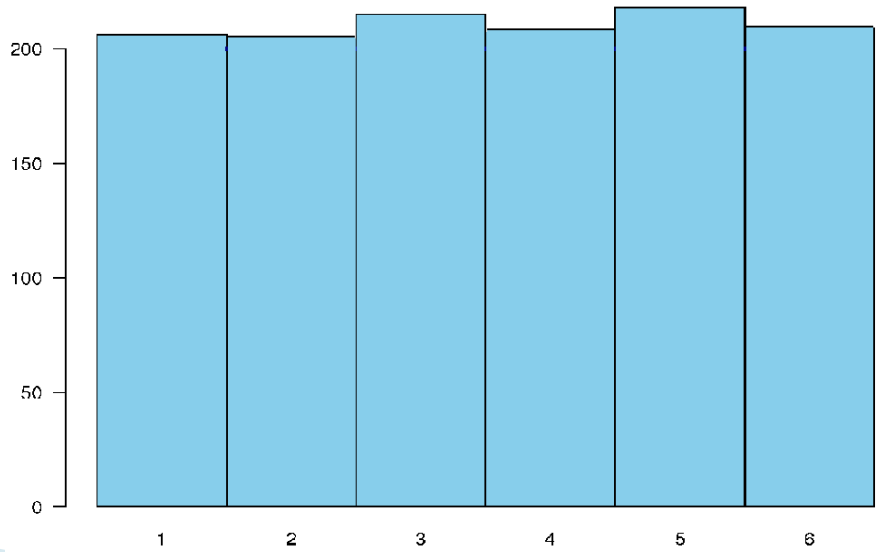
- In the previous example, x is a scalar.
- If we use more than one feature, x becomes a vector.
- If we are using d features, **x** is in a *d*–dimensional Euclidean space R^d , called the *feature space*
- Multivariate Gaussian

Multiclass classification

- In the previous example, it only has to classify instances into one of two classes (binary classification).
- In real-life, there are much more classes, e.g. what is the animal in the picture.
- $\underset{j}{\operatorname{argmax}} P(\omega_j|x)$ where $P(\omega_j/x)$ sum up to one for all j

More about sampling

- Think about another case, let x be the number generated from a dice.
- x is a discrete variable.
- Can you guess the underlying PDF?



More about Bayesian learning

- In the view of data generation, we would like to compute $p(\theta|\mathcal{D})$

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int p(\mathcal{D}|\theta)p(\theta)\, d\theta}$$

- Sometimes, the posterior distribution may not follow any convenient distributional form.
  - In our previous example, both the likelihood and prior are selected as Gaussians (conjugate prior), which is not true for many real world problems.
  - In this case, need to use Monte Carlo sampling

Approximate Bayesian Computation

- In some cases, it is even impossible to write down the likelihood expression.
- Approximate Bayesian Computation (ABC) is proposed to skip the computation of likelihood.
- This is still a hot research topic nowadays.

Reading list

- Pattern Classification, Chapter 1- 3