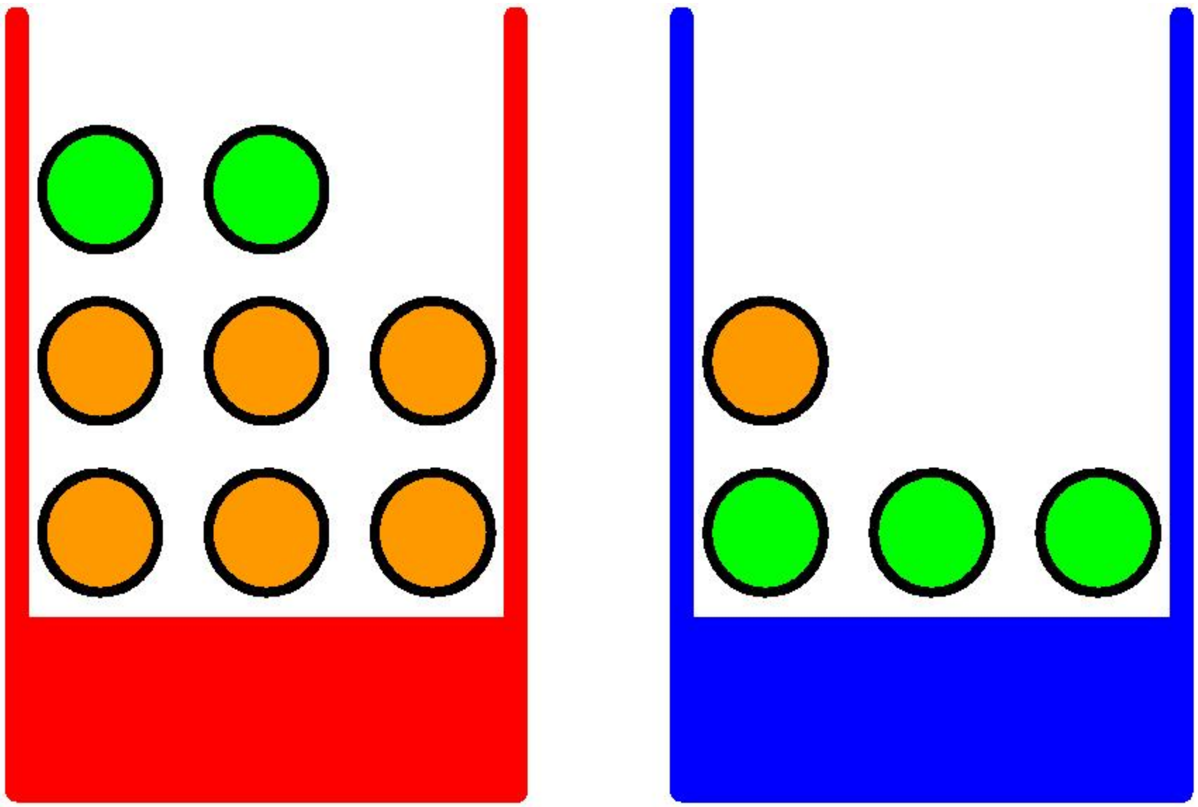
Probability Distributions

Jian Tang
tangjianpku@gmail.com
HEC MONTRÉAL
Institut des algorithmes d'apprentissage de Montréal MILA

Probability Theory

Two boxes with Apples and Oranges



Probability Theory

- (1) Suppose we randomly pick one of the boxes
- (2) Randomly select a fruit from the box
- (3) Observe the type of fruit, and then put it back to where it came from
- Suppose we pick the red box 40% of the time, and the blue box 60 % of the time
- We are equally likely to select any fruit in the boxes

Probability Theory

- Two random variables
  - The identity of the selected box B (B can be red or blue)
  - The identity of the fruit F (F can be apple or orange)
- Define the probability
  - P(B =red) = 4/10, P(B= blue) = 6/10
- Questions:
  - What is the overall probability that the selection procedure will pick an apple, i.e., P(F=apple)=?
  - Given that we have chosen an orange, what is the probability that the box was the blue one, i.e. .P(B=blue|F=orange)?

Two Random Variables

- X: takes the values, x1, x2, …, xm (m =5)
- Y: takes the values, y1, y2, …, yn (n =3)
- n_{ij}: the number of instances x=xi and y=yj
- N: total number of instances

- Joint Probability

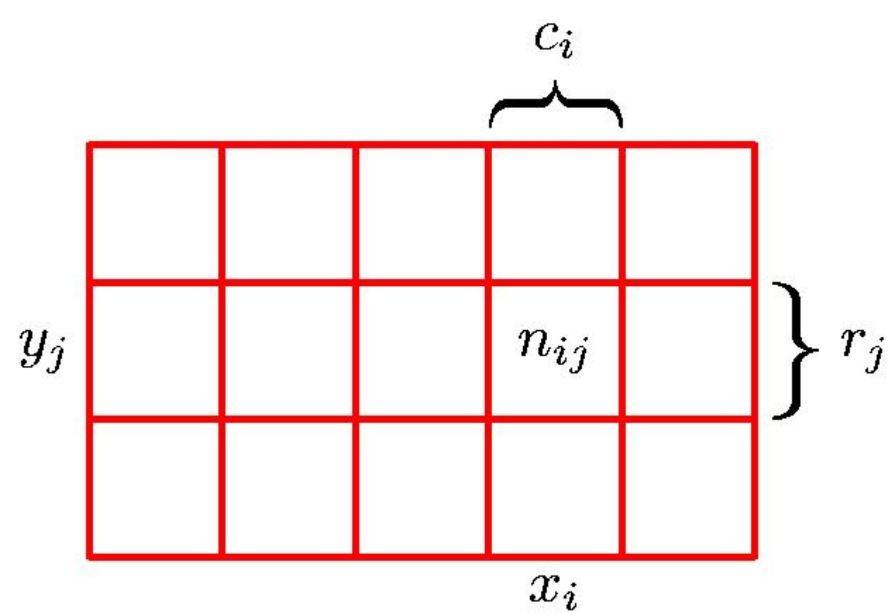
$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

- Marginal Probability

$$p(X = x_i) = \frac{c_i}{N}.$$

- Conditional Probability

$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$

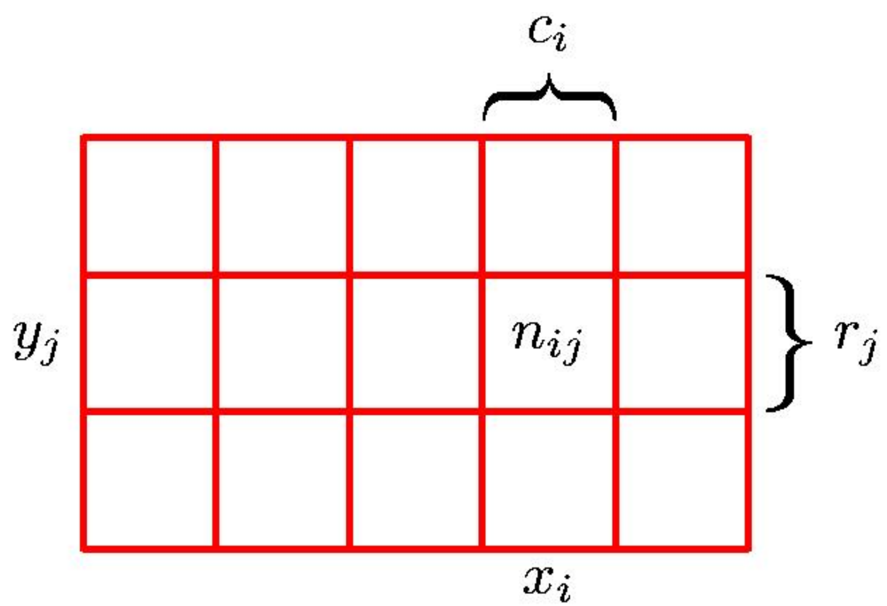


Probability Theory

- Sum Rule

$$p(X = x_i) = \frac{c_i}{N} = \frac{1}{N}\sum_{j=1}^{L} n_{ij}$$

$$= \sum_{j=1}^{L} p(X = x_i, Y = y_j)$$



- Product Rule

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N}$$

$$= p(Y = y_j | X = x_i) p(X = x_i)$$

The Rules of Probability

- Sum Rule $$p(X) = \sum_{Y} p(X, Y)$$
- Product Rule $$p(X, Y) = p(Y|X)p(X)$$

Bayes' Theorem

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

$$p(X) = \sum_Y p(X|Y)p(Y)$$

posterior $\propto$ likelihood × prior

The Fruit Example

- The probabilities of selecting either the red or the blue box:
  - $P(B = red) = 4/10$
  - $P(B = blue) = 6/10$
- Further define the conditional probability
  - $P(F = apple | B = red) = 1/4$
  - $P(F = orange | B = red) = 3/4$
  - $P(F = apple | B = blue) = 3/4$
  - $P(F = orange | B = blue) = 1/4$
- Answers to the questions

$P(F= apple) = P(F=apple|B=red)P(B=red) + P(F=apple|B=blue)P(B=blue)$
$= 1/4 \times 4/10 + 3/4 \times 6/10 = 11/20$

$P(B = red |F=orange) = P(F=orange| B=red) P(B=red)/$
$P(F=orange) = 3/4 \times 4/10 \times 20/9 = 2/3$

Expectations

- Expectations E[f]: the average value of some function f(x) under a probability distribution p(x)

$$\mathbb{E}[f] = \sum_x p(x) f(x)$$

$$\mathbb{E}[f] = \int p(x) f(x) \, \mathrm{d}x$$

$$\mathbb{E}_x[f|y] = \sum_x p(x|y) f(x)$$

Conditional Expectation (discrete)

$$\mathbb{E}[f] \simeq \frac{1}{N} \sum_{n=1}^{N} f(x_n)$$

Approximate Expectation
(discrete and continuous)

Variances and Covariances

- Variances var[f]: a measure of how much variability there is in f(x) around its mean value E[f(x)]

$$\mathrm{var}[f] = \mathbb{E}\left[\left(f(x) - \mathbb{E}[f(x)]\right)^2\right] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2$$

- Covariance of two random variables x and y, cov[x,y]: the extent to which x and y vary together

$$\begin{aligned}\mathrm{cov}[x, y] &= \mathbb{E}_{x,y}\left[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}\right] \\ &= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y]\end{aligned}$$

$$\begin{aligned}\mathrm{cov}[\mathbf{x}, \mathbf{y}] &= \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\{\mathbf{x} - \mathbb{E}[\mathbf{x}]\}\{\mathbf{y}^{\mathrm{T}} - \mathbb{E}[\mathbf{y}^{\mathrm{T}}]\}\right] \\ &= \mathbb{E}_{\mathbf{x},\mathbf{y}}[\mathbf{x}\mathbf{y}^{\mathrm{T}}] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}^{\mathrm{T}}]\end{aligned}$$

Binomial Distribution

- A Binary variable $x \in \{0, 1\}$, e.g., Flipping a coin. X = 1 representing heads and X = 0 representing tails. Define the probability of obtaining heads as:

$$P(X=1) = u$$

- The distribution of the number **m** of observations of x=1 (e.g. the number of heads).
- The probability of observing m heads given N coin flips and a parameter μ is given by:

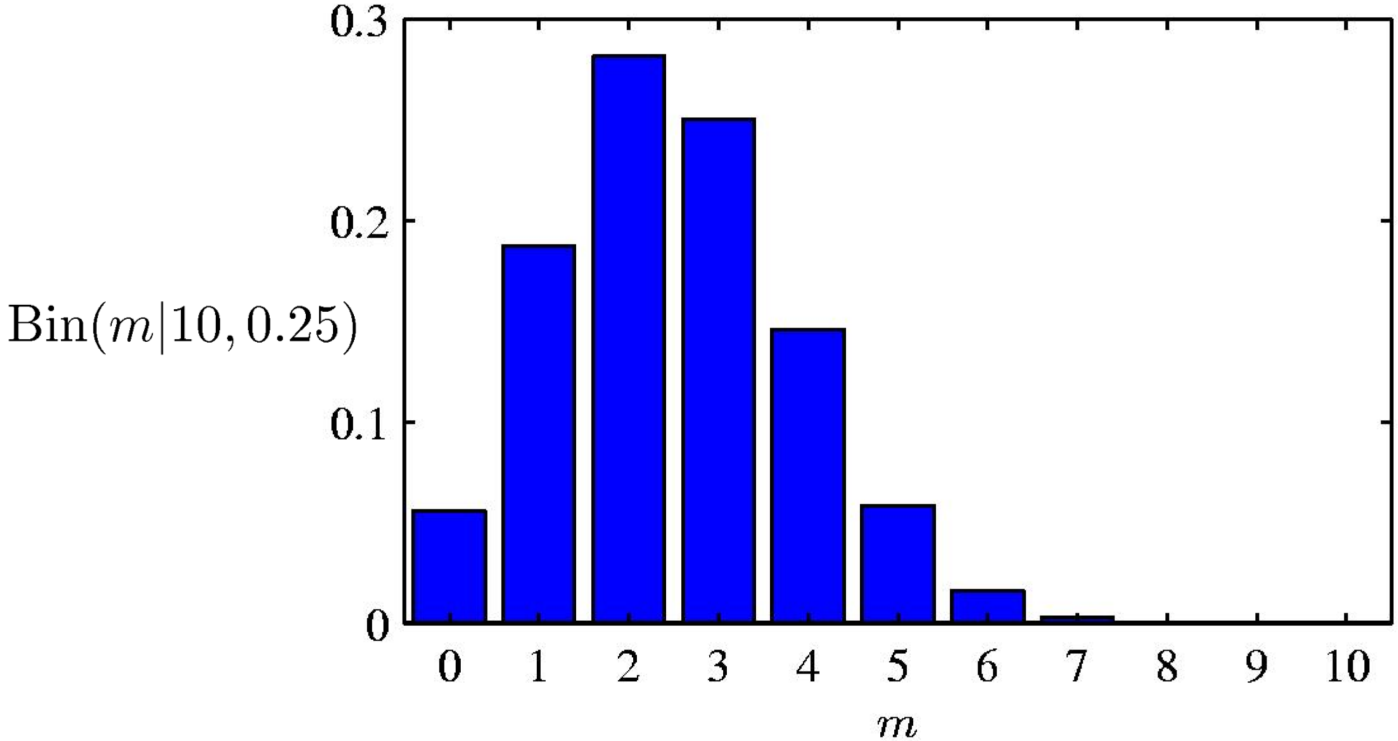
$$p(m \text{ heads}|N, \mu) =$$
$$\mathrm{Bin}(m|N, \mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m}$$

- The mean and variance can be easily derived as:

$$\mathbb{E}[m] \equiv \sum_{m=0}^{N} m\mathrm{Bin}(m|N, \mu) = N\mu$$
$$\mathrm{var}[m] \equiv \sum_{m=0}^{N} (m - \mathbb{E}[m])^2 \, \mathrm{Bin}(m|N, \mu) = N\mu(1-\mu)$$

Example

•Histogram plot of the Binomial distribution as a function of m for N=10 and μ =0.25.



Multinomial Variables

- Consider a random variable that can take on one of K possible mutually exclusive states (e.g. roll of a dice).

- We will use so-called 1-of-K encoding scheme.

- If a random variable can take on K=6 states, and a particular observation of the variable corresponds to the state $x_3=1$, then **x** will be resented as:

1-of-K coding scheme: $$\mathbf{x} = (0, 0, 1, 0, 0, 0)^{\mathrm{T}}$$

- If we denote the probability of $x_k=1$ by the parameter μ_k, then the distribution over **x** is defined as:

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k} \qquad \forall k : \mu_k \geqslant 0 \quad \text{and} \quad \sum_{k=1}^{K} \mu_k = 1$$

Multinomial Variables

- Multinomial distribution can be viewed as a generalization of Bernoulli distribution to more than two outcomes.

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k}$$

- It is easy to see that the distribution is normalized:

$$\sum_{\mathbf{x}} p(\mathbf{x}|\boldsymbol{\mu}) = \sum_{k=1}^{K} \mu_k = 1$$

and

$$\mathbb{E}[\mathbf{x}|\boldsymbol{\mu}] = \sum_{\mathbf{x}} p(\mathbf{x}|\boldsymbol{\mu})\mathbf{x} = (\mu_1, \ldots, \mu_K)^{\mathrm{T}} = \boldsymbol{\mu}$$

Maximum Likelihood Estimation

- Suppose we observed a dataset $\mathcal{D} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$
- We can construct the likelihood function, which is a function of μ.

$$p(\mathcal{D}|\boldsymbol{\mu}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mu_k^{x_{nk}} = \prod_{k=1}^{K} \mu_k^{(\sum_n x_{nk})} = \prod_{k=1}^{K} \mu_k^{m_k}$$

- Note that the likelihood function depends on the N data points only though the following K quantities:

$$m_k = \sum_n x_{nk}, \quad k = 1, ..., K.$$

which represents the number of observations of x_k=1.

- These are called the sufficient statistics for this distribution.

Maximum Likelihood Estimation

$$p(\mathcal{D}|\boldsymbol{\mu}) = \prod_{n=1}^{N}\prod_{k=1}^{K}\mu_k^{x_{nk}} = \prod_{k=1}^{K}\mu_k^{(\sum_n x_{nk})} = \prod_{k=1}^{K}\mu_k^{m_k}$$

- To find a maximum likelihood solution for μ, we need to maximize the log-likelihood taking into account the constraint that $\sum_k \mu_k = 1$

- Forming the Lagrangian:

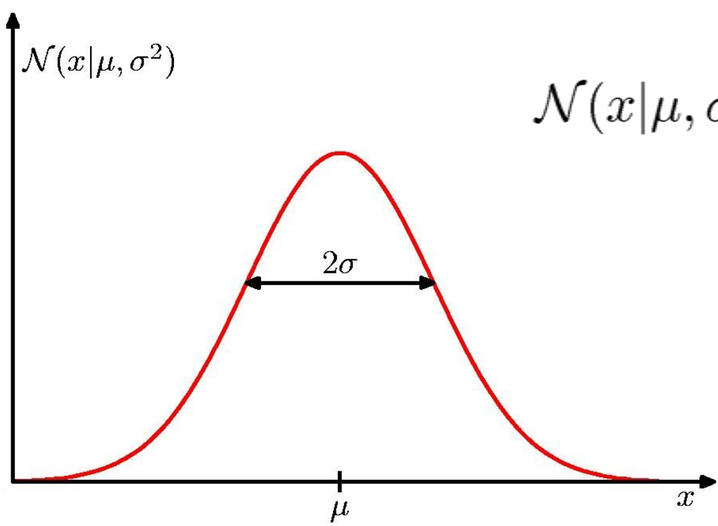
$$\sum_{k=1}^{K} m_k \ln \mu_k + \lambda\left(\sum_{k=1}^{K}\mu_k - 1\right)$$

$$\mu_k = -m_k/\lambda \qquad \mu_k^{\mathrm{ML}} = \frac{m_k}{N} \qquad \lambda = -N$$

which is the fraction of observations for which x_k=1.

Gaussian Univariate Distribution

- In the case of a single variable x, the Gaussian distribution takes form:



$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

which is governed by two parameters:

- *μ* (mean)
- *σ*↑2 (variance)

- The Gaussian distribution satisfies:

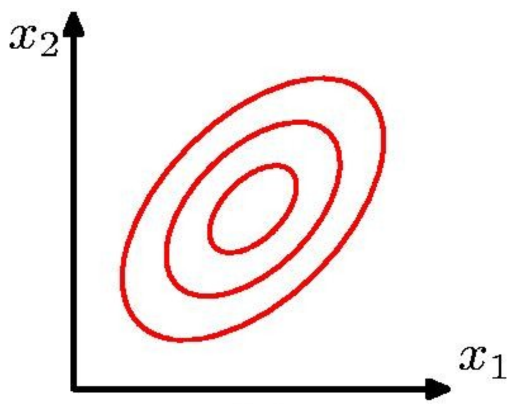
$$\mathcal{N}(x|\mu,\sigma^2) > 0$$

$$\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) \,\mathrm{d}x = 1$$

Multivariate Gaussian Distribution

- For a D-dimensional vector **x**, the Gaussian distribution takes form:

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$



which is governed by two parameters:

- μ is a D-dimensional mean vector.
- Σ is a D by D covariance matrix.

and $|\Sigma|$ denotes the determinant of Σ.

- Note that the covariance matrix is a symmetric positive definite matrix.

Maximum Likelihood Estimation

- Suppose we observed i.i.d data $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}.$

- We can construct the log-likelihood function, which is a function of μ and § :

$$\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu})$$

- Note that the likelihood function depends on the N data points only though the following sums:

Sufficient Statistics

$$\sum_{n=1}^{N}\mathbf{x}_n \qquad \sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^{\mathrm{T}}$$

Maximum Likelihood Estimation

•To find a maximum likelihood estimate of the mean, we set the derivative of the log-likelihood function to zero:

$$\frac{\partial}{\partial \boldsymbol{\mu}} \ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}) = 0$$

and solve to obtain:

$$\boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n.$$

• Similarly, we can find the ML estimate of Σ :

$$\boldsymbol{\Sigma}_{\mathrm{ML}} = \frac{1}{N} \sum_{n=1}^{N} (\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}}.$$

Maximum Likelihood Estimation

- Evaluating the expectation of the ML estimates under the true distribution, we obtain:

Unbiased estimate

$$\begin{aligned} \mathbb{E}[\boldsymbol{\mu}_{\mathrm{ML}}] &= \boldsymbol{\mu} \\ \mathbb{E}[\boldsymbol{\Sigma}_{\mathrm{ML}}] &= \frac{N-1}{N}\boldsymbol{\Sigma}. \end{aligned}$$

Biased estimate

- Note that the maximum likelihood estimate of Σ is biased.

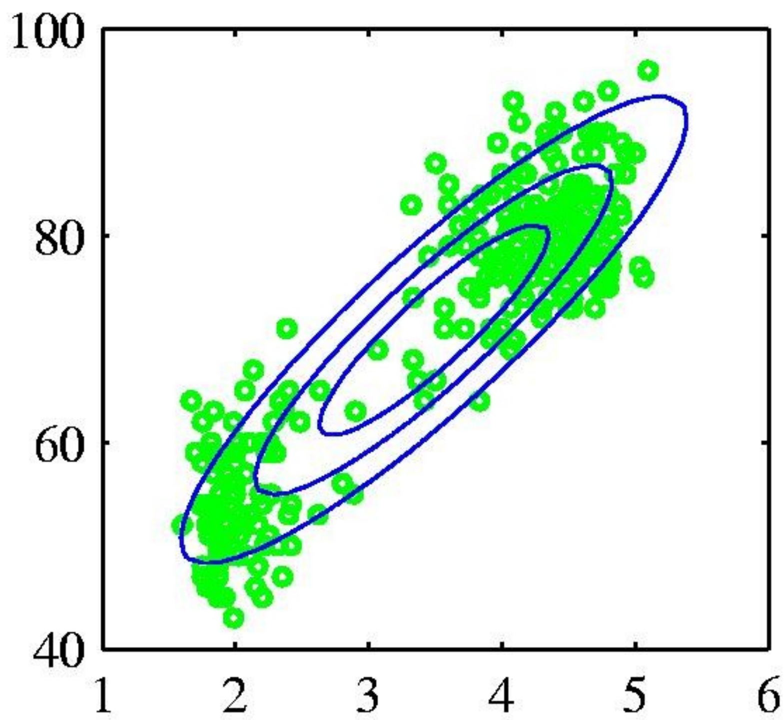
- We can correct the bias by defining a different estimator:

$$\widetilde{\boldsymbol{\Sigma}} = \frac{1}{N-1}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}}.$$

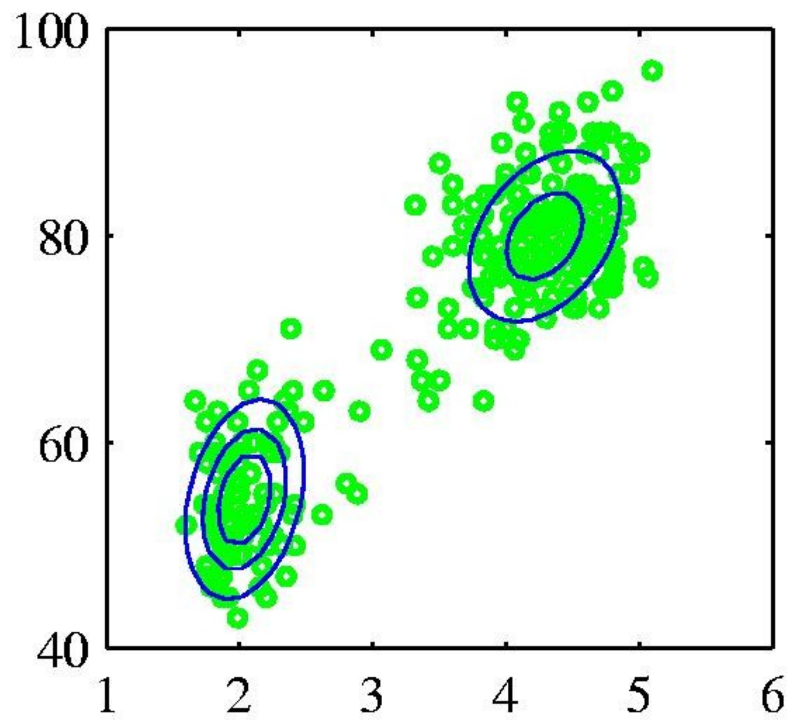
Mixture of Gaussians

•When modeling real-world data, Gaussian assumption may not be appropriate.

• Consider the following example: Old Faithful Dataset



Single Gaussian



Mixture of two Gaussians

Mixture of Gaussians

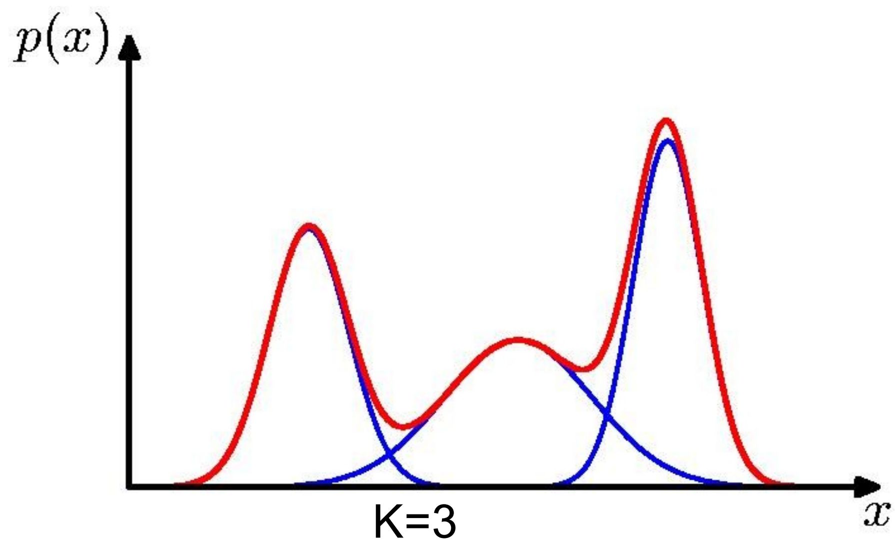
•We can combine simple models into a complex model by defining a superposition of K Gaussian densities of the form:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Component

Mixing coefficient

$$\forall k : \pi_k \geqslant 0 \qquad \sum_{k=1}^{K} \pi_k = 1$$

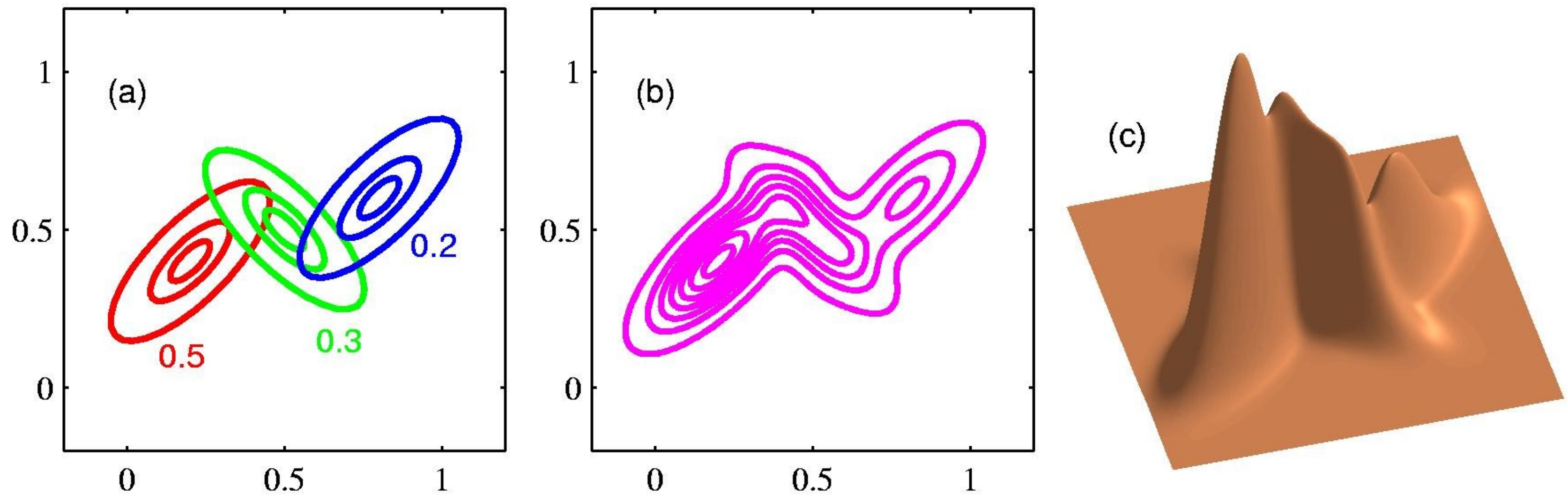


•Note that each Gaussian component has its own mean μ_k and covariance ${}_k$. The parameters $\pi \downarrow k$ are called mixing coefficients.

•Mote generally, mixture models can comprise linear combinations of other distributions.

Mixture of Gaussians

- Illustration of a mixture of 3 Gaussians in a 2-dimensional space:



(a)Contours of constant density of each of the mixture components, along with the mixing coefficients

(b) Contours of marginal probability density $p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$

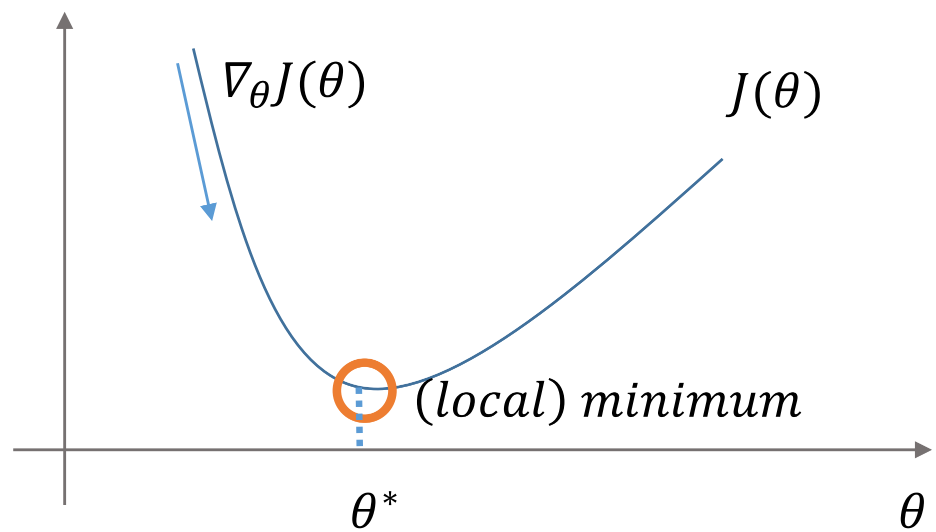
(c) A surface plot of the distribution p(x).

Gradient Descent

- Gradient descent is a way to **minimize** an objective function $J(\theta)$
  - $J(\theta)$: objective function
  - $\theta \in R^{d}$: model's parameters
  - η: learning rate, which determines the size of the steps we take to reach a (local) minimum.

Update equation

$$\theta = \theta - \eta * \nabla_{\theta} J(\theta)$$



Slides from St_Hakky

References

- Chap. 1&2, Bishop, Patten Recognition and Machine Learning.