STA414/2104

Statistical Methods for Machine Learning II

Murat A. Erdogdu

Department of Computer Science
Department of Statistical Sciences

Lecture 2

UNIVERSITY OF TORONTO

Announcements

- Midterm is “in class’’ 2 hr long written exam, to be held on March 1st for Mon section, and March 2nd for Tue section.
- HW1 will be released next week.
- TA office hours for HW1 will be posted.

Last Time

- Supervised vs unsupervised learning
- Least squares
- Polynomial curve fitting
- Overfitting and generalization
- Effect of regularization
- Cross validation

Maximum Likelihood Estimation (MLE)

- **MLE** is a method to estimate the unknown parameter by maximizing the likelihood function, so the observed data is most probable.

- **Recipe:**
  - Observe data: $\mathcal{D} = \{x_1, x_2, ..., x_N\}$
  - Assume data is iid from a distribution $x_i \sim p(x|\theta)$
  - Write down the joint density

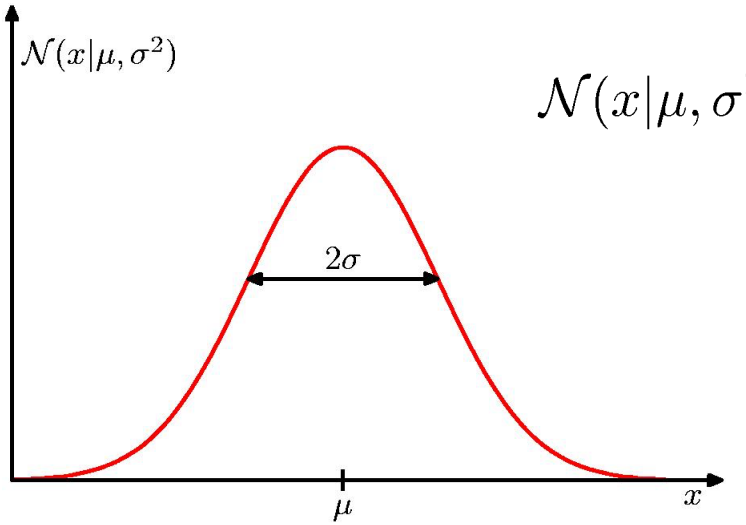
$$p(x_1, x_2, ..., x_N|\theta) = \prod_{i=1}^{N} p(x_i|\theta) \quad = \mathcal{L}(\theta; x_1, x_2, ..., x_N)$$

  - Plug in the observed values (data) and see it as a function of the unknown parameters (at this stage, we call this function the likelihood)
  - Maximize the likelihood. $\hat{\theta}^{\mathrm{ML}} = \arg\max_{\theta} \mathcal{L}(\theta; x_1, x_2, ..., x_N)$

- We generally minimize negative log-likelihood since log is monotone strictly increasing function, and converts products to summations (which behave nicely taking derivatives). See next example for a demonstration.

Univariate Gaussian 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:

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

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

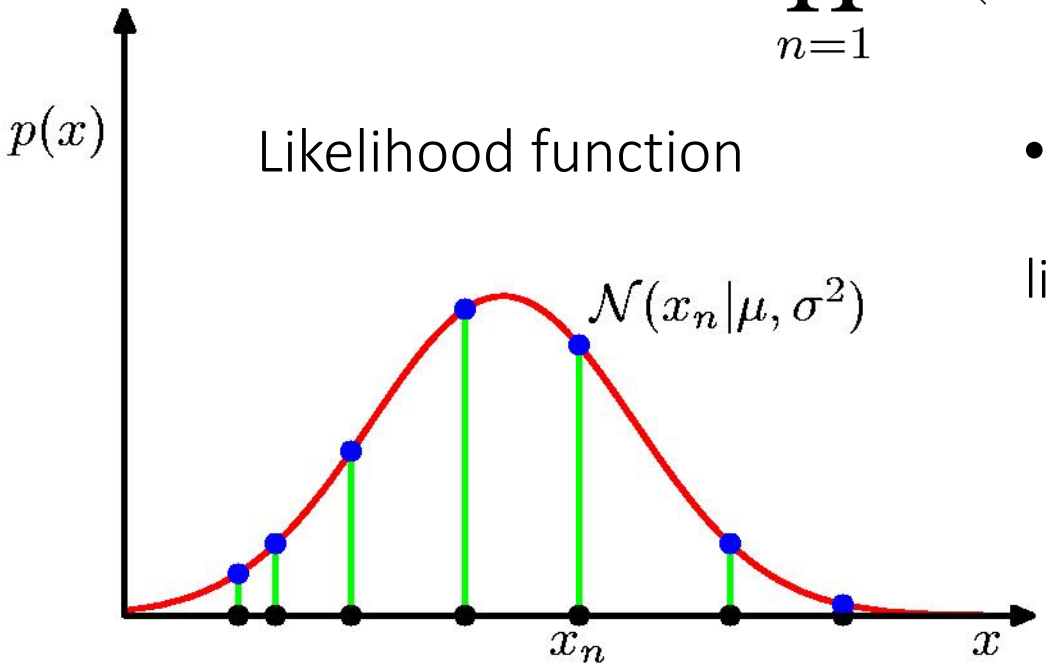
Gaussian Parameter Estimation

• Suppose we have a dataset of i.i.d. observations $\mathbf{x} = (x_1,\ldots,x_N)^T$, representing N 1-dimensional observations.

• Because our data x is i.i.d., we can write down the joint probability of all the data points as given μ and σ^2 :

$$p(\mathbf{x}|\mu,\sigma^2) = \prod_{n=1}^{N} \mathcal{N}\left(x_n|\mu,\sigma^2\right) = \mathcal{L}(\mu,\sigma^2;\mathbf{x})$$

Likelihood function



• When viewed as a function of μ and σ^2 , this is called the likelihood function.

Maximum (log) likelihood

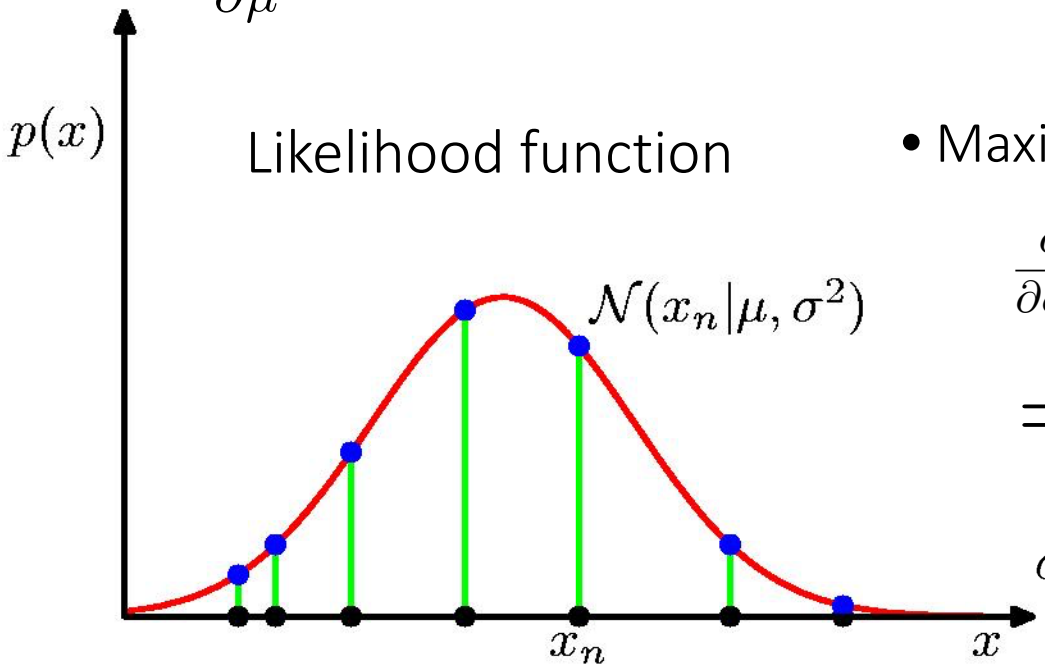
- The log-likelihood can be written as:

$$\ln p\left(\mathbf{x}|\mu,\sigma^2\right) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)$$

- Maximizing w.r.t. μ gives:

Sample mean

$$\frac{\partial}{\partial\mu}\ln p(\mathbf{x}|\mu,\sigma^2) = 0 \quad \Longrightarrow \quad \mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n$$



Likelihood function

- Maximizing w.r.t σ^2 gives:

$$\frac{\partial}{\partial\sigma^2}\ln p(\mathbf{x}|\mu_{\mathrm{ML}},\sigma^2) = 0$$

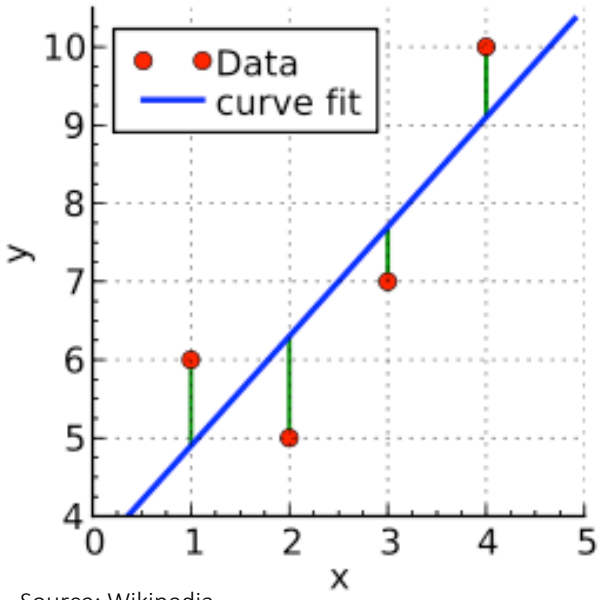
$$\Longrightarrow$$

Sample variance

$$\sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2$$

Linear Least Squares

From last class: Minimize the sum of the squares of the errors between the predictions $y(\mathbf{x}_n, \mathbf{w})$ for each data point x_n and the corresponding real-valued targets t_n.



Source: Wikipedia

Loss function: sum-of-squared error function:

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n^T\mathbf{w} - t_n)^2$$

$$= \frac{1}{2}(\mathbf{X}\mathbf{w} - \mathbf{t})^{\mathbf{T}}(\mathbf{X}\mathbf{w} - \mathbf{t}).$$

$$= \tfrac{1}{2}\mathbf{w}^T(\mathbf{X}^T\mathbf{X})\mathbf{w} - (\mathbf{X}^T\mathbf{t})^T\mathbf{w} + \tfrac{1}{2}\|\mathbf{t}\|^2$$

$$\text{minimize} E(\mathbf{w}) = \text{solve} \ \{\nabla_{\mathbf{w}} E(\mathbf{w}) = 0\}$$

Notation:

$$\nabla_{\mathbf{w}} f(\mathbf{w}) = \frac{\partial f(\mathbf{w})}{\partial \mathbf{w}} = \begin{bmatrix} \frac{df(\mathbf{w})}{d\mathbf{w}_1} \\ \vdots \\ \frac{df(\mathbf{w})}{d\mathbf{w}_d} \end{bmatrix}$$

For symmetric matrix $\mathbf{A}$ and a vector $\mathbf{a}$

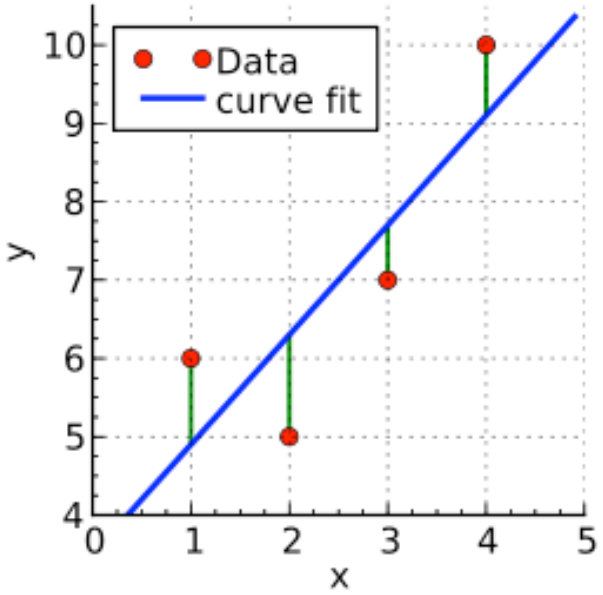
$$\nabla_{\mathbf{w}}(\tfrac{1}{2}\mathbf{w}^T\mathbf{A}\mathbf{w}) = \mathbf{A}\mathbf{w} \quad \nabla_{\mathbf{w}}(\mathbf{a}^T\mathbf{w}) = \mathbf{a}$$

Linear Least Squares

If $\mathbf{X^T X}$ is nonsingular, then the unique solution is given by:

$$\nabla_{\mathbf{w}} E(\mathbf{w}) = \frac{\partial E(\mathbf{w})}{\partial \mathbf{w}}$$

$$\nabla_{\mathbf{w}} E(\mathbf{w}) = \mathbf{X}^T\mathbf{X}\mathbf{w} - \mathbf{X}^T\mathbf{t} = 0$$



Source: Wikipedia

optimal weights

vector of target values

$$\mathbf{w}^* = (\mathbf{X^T X})^{-1}\mathbf{X^T t}$$

the design matrix has one input vector per row

- At an arbitrary input $\mathbf{x}_0$ the prediction is $y(\mathbf{x}_0, \mathbf{w}) = \mathbf{x}_0^T \mathbf{w}^*$.
- The entire model is characterized by d+1 parameters $\mathbf{w}^*$.

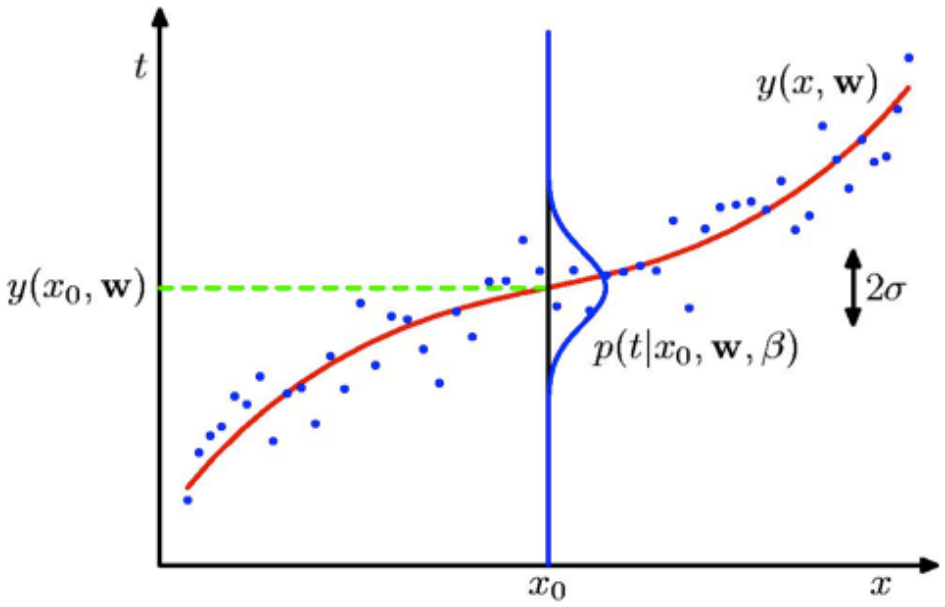
Probabilistic Perspective

• We saw that polynomial curve fitting can be expressed in terms of error minimization. We now view it from probabilistic perspective.

• Suppose that our model arose from a statistical model:

$$t = y(\mathbf{x}, \mathbf{w}) + \epsilon,$$

where ϵ is a random error having Gaussian distribution with zero mean, and is independent of x.



Thus we have:

$$p(t|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(t|y(\mathbf{x}, \mathbf{w}), \beta^{-1}),$$

where β is a precision parameter, corresponding to the inverse variance.

Maximum Likelihood

If the data are assumed to be independently ~~and identically~~ distributed ~~(i.i.d assumption)~~, the likelihood function takes form:

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|y(\mathbf{x}_n, \mathbf{w}), \beta^{-1}).$$

It is often convenient to maximize the log of the likelihood function: Independence assumption makes the log-likelihood a sum, so its derivative can be taken term by term.

$$\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \underbrace{-\frac{\beta}{2} \sum_{n=1}^{N} (y(\mathbf{x}_n, \mathbf{w}) - t_n)^2}_{\beta E(\mathbf{w})} + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi).$$

- Maximizing log-likelihood with respect to w (under the assumption of a Gaussian noise) is equivalent to minimizing the sum-of-squared error function.

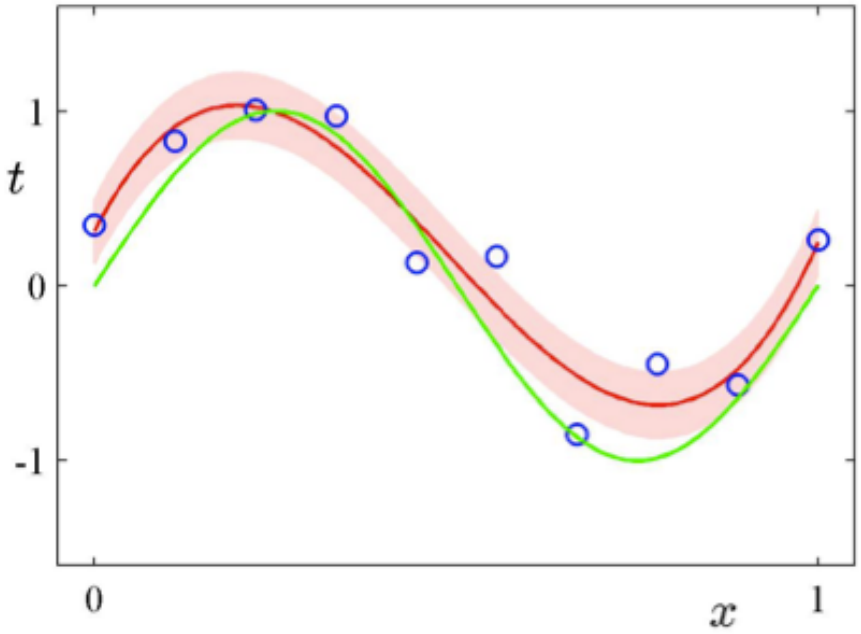
- Determine $\mathbf{w}_{ML}$ by maximizing log-likelihood. Then maximizing w.r.t. β

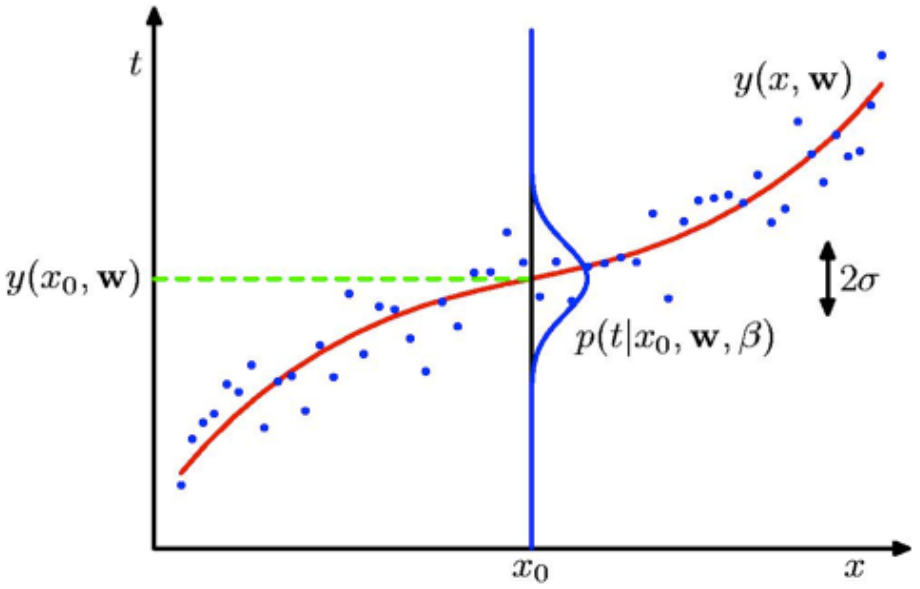
$$\frac{1}{\beta_{ML}} = \frac{1}{N} \sum_{n} (y(\mathbf{x}_n, \mathbf{w}_{ML}) - t_n)^2.$$

Predictive Distribution

Once we determined the parameters w and β, we can make prediction for new values of x:

$$p(t|\mathbf{x}, \mathbf{w}_{ML}, \beta_{ML}) = \mathcal{N}(t|y(\mathbf{x}, \mathbf{w}_{ML}), \beta_{ML}^{-1}).$$





Linear Basis Function Models

- Remember, the simplest linear model for regression:

$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + w_2 x_2 + ... + w_d x_d = w_0 + \sum_{j=1}^{d} w_j x_j,$$

where $\mathbf{x} = (x_1, x_2, ..., x_d)^T$ a d-dimensional input vector (covariates).

Key property: linear function of the parameters $w_0, w_1, ..., w_d$

- However, it is also a linear function of input variables.
  Instead consider:

$$y(\mathbf{x}, \mathbf{w}) = w_0 \phi_0(\mathbf{x}) + w_1 \phi_1(\mathbf{x}) + ... + w_{M-1} \phi_{M-1}(\mathbf{x}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}),$$

where $\phi_j(\mathbf{x})$ are known as basis functions.

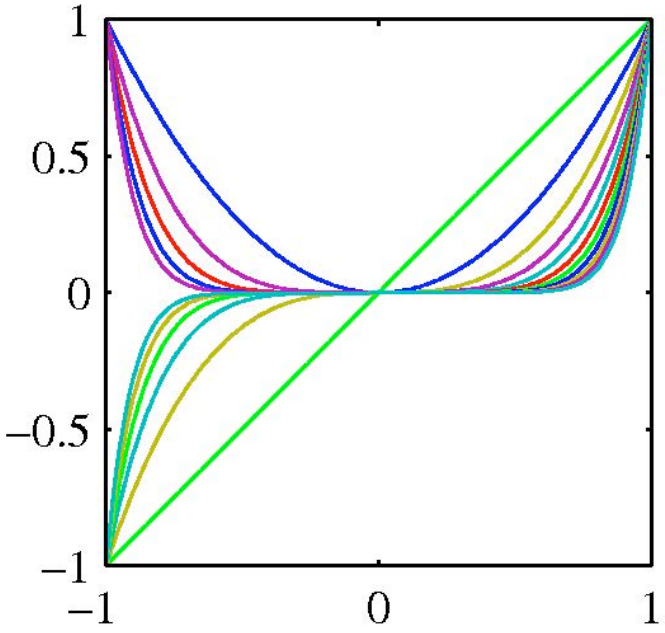
- Typically $\phi_0(\mathbf{x}) = 1$ so that w_0 acts as a bias (or intercept).
- In the simplest case, we use linear bases functions: $\phi_j(\mathbf{x}) = x_j$.
- Using nonlinear basis allows the functions $y(\mathbf{x}, \mathbf{w})$ to be nonlinear functions of the input space.

Linear Basis Function Models

Polynomial basis functions:

$$\phi_j(x) = x^j .$$

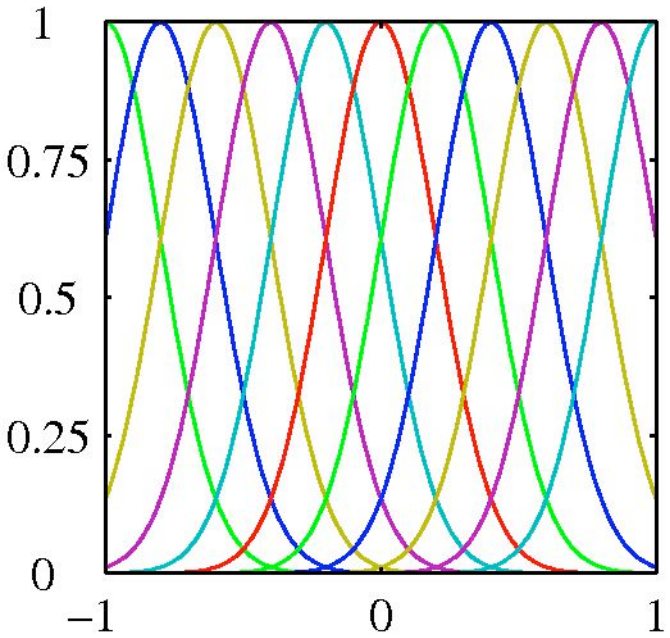
(as in lecture 1)



Basis functions are global: small changes in x affect all basis functions.

Gaussian basis functions:

$$\phi_j(x) = \exp\left(-\frac{(x-\mu_j)^2}{2s^2}\right).$$

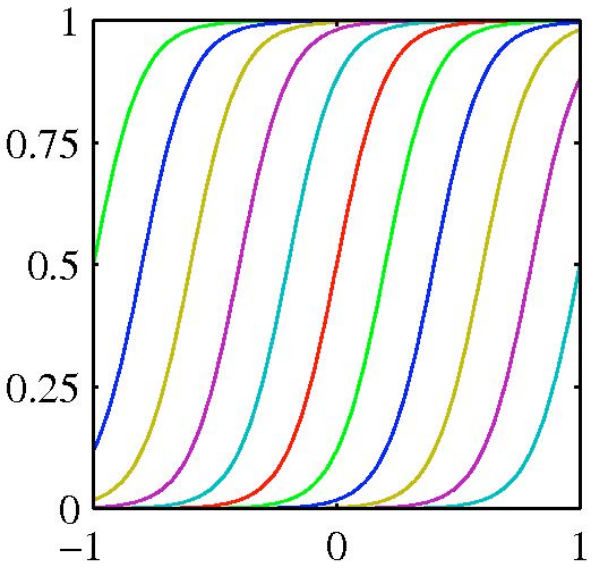


Basis functions are local: small changes in x only affect nearby basis functions.
μ_j and s control location and scale (width).

Linear Basis Function Models

Sigmoidal basis functions

$$\phi_j(x) = \sigma\left(\frac{x - \mu_j}{s}\right), \text{ where } \sigma(a) = \frac{1}{1 + \exp(-a)}.$$



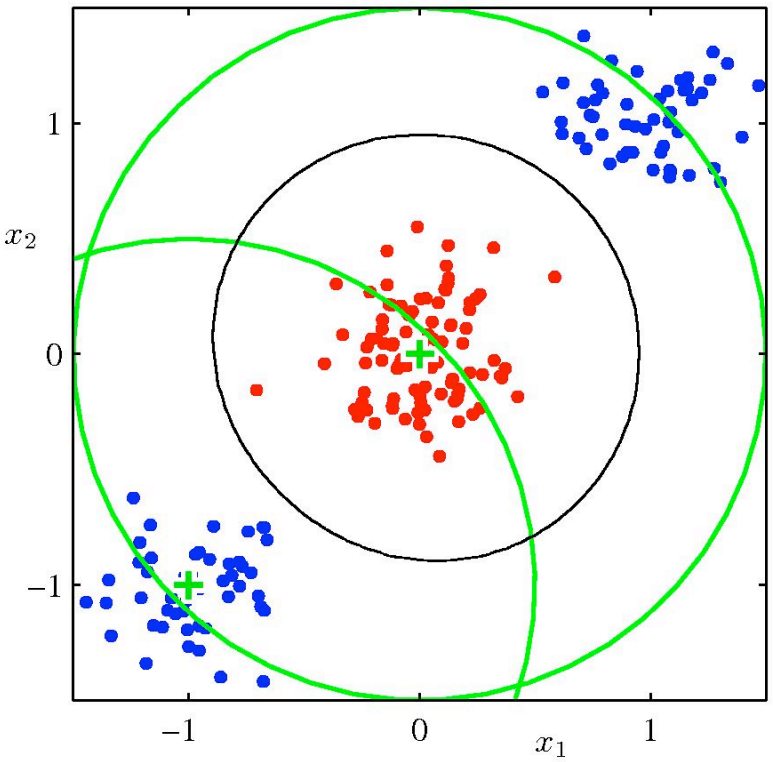
Basis functions are local: small changes in x only affect nearby basis functions.
μ_j and s control location and scale (slope).

• Basis functions can be used for regression and classification. In the case of classification, decision boundaries will be linear in the feature space $\boldsymbol{\phi}$, but would correspond to nonlinear boundaries in the original input space x.

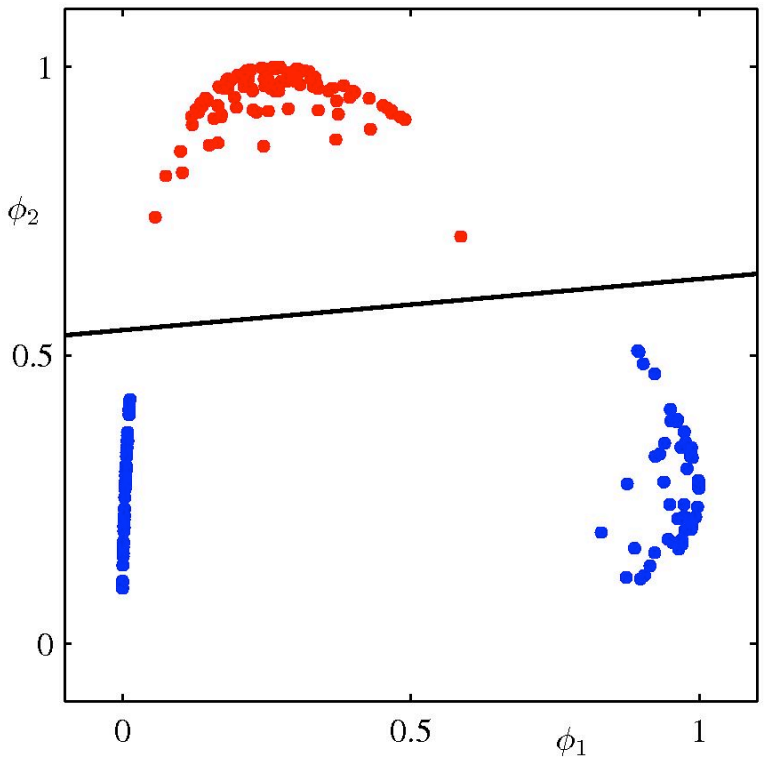
• Classes that are linearly separable in the feature space $\boldsymbol{\phi}(\mathbf{x})$ need not be linearly separable in the original input space.

Linear Basis Function Models

Original input space



Corresponding feature space using two Gaussian basis functions



• We define two Gaussian basis functions with centers shown by the green crosses, and with contours shown by the green circles.

• Linear decision boundary (right) is obtained by using logistic regression (to be covered in the next lecture), and corresponds to the nonlinear decision boundary in the input space (left, black curve).

Maximum Likelihood

• As before, assume observations arise from a deterministic function with an additive Gaussian noise:

$$t = y(\mathbf{x}, \mathbf{w}) + \epsilon,$$

which we can write as:

$$p(t|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(t|y(\mathbf{x}, \mathbf{w}), \beta^{-1}).$$

• Given observed inputs $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$, and corresponding target values $\mathbf{t} = [t_1, t_2, ..., t_N]^T$ under independence assumption, we can write down the likelihood function:

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{i=1}^{N} \mathcal{N}(t_n|\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_n), \beta),$$

where $\boldsymbol{\phi}(\mathbf{x}) = (\phi_0(\mathbf{x}), \phi_1(\mathbf{x}), ..., \phi_{M-1}(\mathbf{x}))^T$.

Maximum Likelihood

Taking the logarithm, we obtain:

$$\ln p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \sum_{i=1}^{N} \ln \mathcal{N}(t_n|\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n), \beta)$$

$$= -\frac{\beta}{2} \underbrace{\sum_{n=1}^{N} \left(t_n - \mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n)\right)^2}_{\text{sum-of-squares error function}} + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi).$$

Differentiating and setting to zero yields:

$$\nabla_{\mathbf{w}} \ln p(\mathbf{t}|\mathbf{w}, \beta) = \beta \sum_{n=1}^{N} \left\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\right\} \boldsymbol{\phi}(\mathbf{x}_n)^{\mathrm{T}} = \mathbf{0}.$$

Maximum Likelihood

Differentiating and setting to zero yields:

$$\nabla_{\mathbf{w}} \ln p(\mathbf{t}|\mathbf{w}, \beta) = \beta \sum_{n=1}^{N} \left\{ t_n - \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n) \right\} \boldsymbol{\phi}(\mathbf{x}_n)^{\mathrm{T}} = \mathbf{0}.$$

Solving for w, we get:

$$\mathbf{w}_{\mathrm{ML}} = \left(\boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi} \right)^{-1} \boldsymbol{\Phi}^{\mathrm{T}} \mathbf{t}$$

Inverse of the matrix (or pseudo-inverse)

where $\boldsymbol{\Phi}$ is known as the design matrix:

$$\boldsymbol{\Phi} = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \cdots & \phi_{M-1}(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \cdots & \phi_{M-1}(\mathbf{x}_N) \end{pmatrix}.$$

Geometry of Least Squares

$$\mathbf{y} = \mathbf{\Phi w}$$

$$= \begin{pmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \cdots & \phi_{M-1}(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \cdots & \phi_{M-1}(\mathbf{x}_N) \end{pmatrix} \cdot \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_{M-1} \end{pmatrix}$$

$$\boldsymbol{\varphi}_0 \quad \boldsymbol{\varphi}_1$$

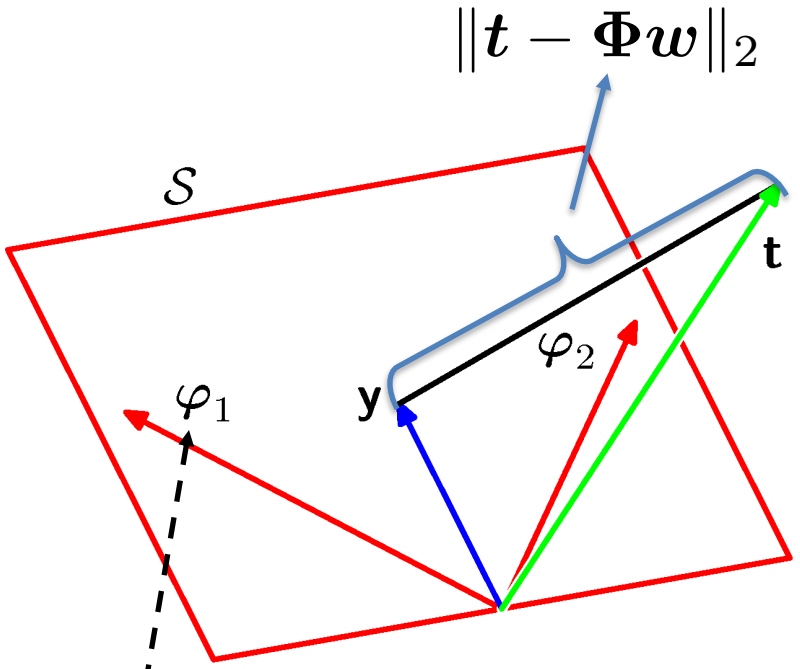
$$= \sum_{m=0}^{M-1} \boldsymbol{\varphi}_m w_m$$

- This means that $\mathbf{y} = \mathbf{\Phi w}$ is in the column space of $\mathbf{\Phi}$
- Consider an N-dimensional space, so that $\mathbf{t} = [t_1, t_2, ..., t_N]^T$ is a vector in that space.

We are looking for **w** so that our prediction **y** (which is always in the column space) is close to **t.**

Geometry of Least Squares

- If M is less than N, then the M basis function $\phi_j(\mathbf{x}_n)$, will span a linear subspace S of dimensionality M.
- S is the column space of $\mathbf{\Phi}$
- Define: $\mathbf{y} = \mathbf{\Phi}\mathbf{w}_{\mathbf{ML}}$.



- The sum-of-squares error is equal to the squared Euclidean distance between **y** and **t** (up to a factor of 1/2).

$$E(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{t} - \mathbf{\Phi}\boldsymbol{w}\|_2^2$$

$$\mathbf{\Phi} = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \cdots & \phi_{M-1}(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \cdots & \phi_{M-1}(\mathbf{x}_N) \end{pmatrix}.$$

The solution corresponds to the orthogonal projection of **t** onto the subspace S.

Bernoulli Distribution

• Consider a single binary random variable $x \in \{0, 1\}$. For example, x can describe the outcome of flipping a coin:

Coin flipping: heads = 1, tails = 0.

• The probability of x=1 will be denoted by the parameter μ, so that:

$$p(x = 1|\mu) = \mu \qquad 0 \leq \mu \leq 1.$$

• The probability distribution, known as Bernoulli distribution, can be written as:

$$\begin{aligned} p(x|\mu) = \mathrm{Bern}(x|\mu) &= \mu^x(1-\mu)^{1-x} \\ \mathbb{E}[x] &= \mu \\ \mathrm{var}[x] &= \mu(1-\mu) \end{aligned}$$

Probability mass function (pmf)

Verify!

Parameter Estimation

- Suppose we observed a dataset $\mathcal{D} = \{x_1, ..., x_N\}$
- We can construct the likelihood function, which is a function of μ.

$$p(\mathcal{D}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}$$

- Equivalently, we can maximize the log of the likelihood function:

$$\ln p(\mathcal{D}|\mu) = \sum_{n=1}^{N} \ln p(x_n|\mu) = \sum_{n=1}^{N} \{x_n \ln \mu + (1-x_n)\ln(1-\mu)\}$$

- Note that the likelihood function depends on the N observations x_n only through the sum $\sum_n x_n$

Sufficient Statistic

Parameter depends on data only through sufficient statistics.

Parameter Estimation

- Suppose we observed a dataset $\mathcal{D} = \{x_1, ..., x_N\}$
- Log-likelihood is

$$\ln p(\mathcal{D}|\mu) = \sum_{n=1}^{N} \ln p(x_n|\mu) = \sum_{n=1}^{N} \{x_n \ln \mu + (1 - x_n) \ln(1 - \mu)\}$$

- Setting the derivative of the log-likelihood function w.r.t μ to zero, we obtain:

$$\mu_{\mathrm{ML}} = \frac{1}{N} \sum_{n=1}^{N} x_n = \frac{m}{N}$$

where m is the number of heads.

Binomial Distribution

• We can also work out 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:

$$\mathrm{Bin}(m|N,\mu) :$$

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

• The mean and variance can be easily derived as:

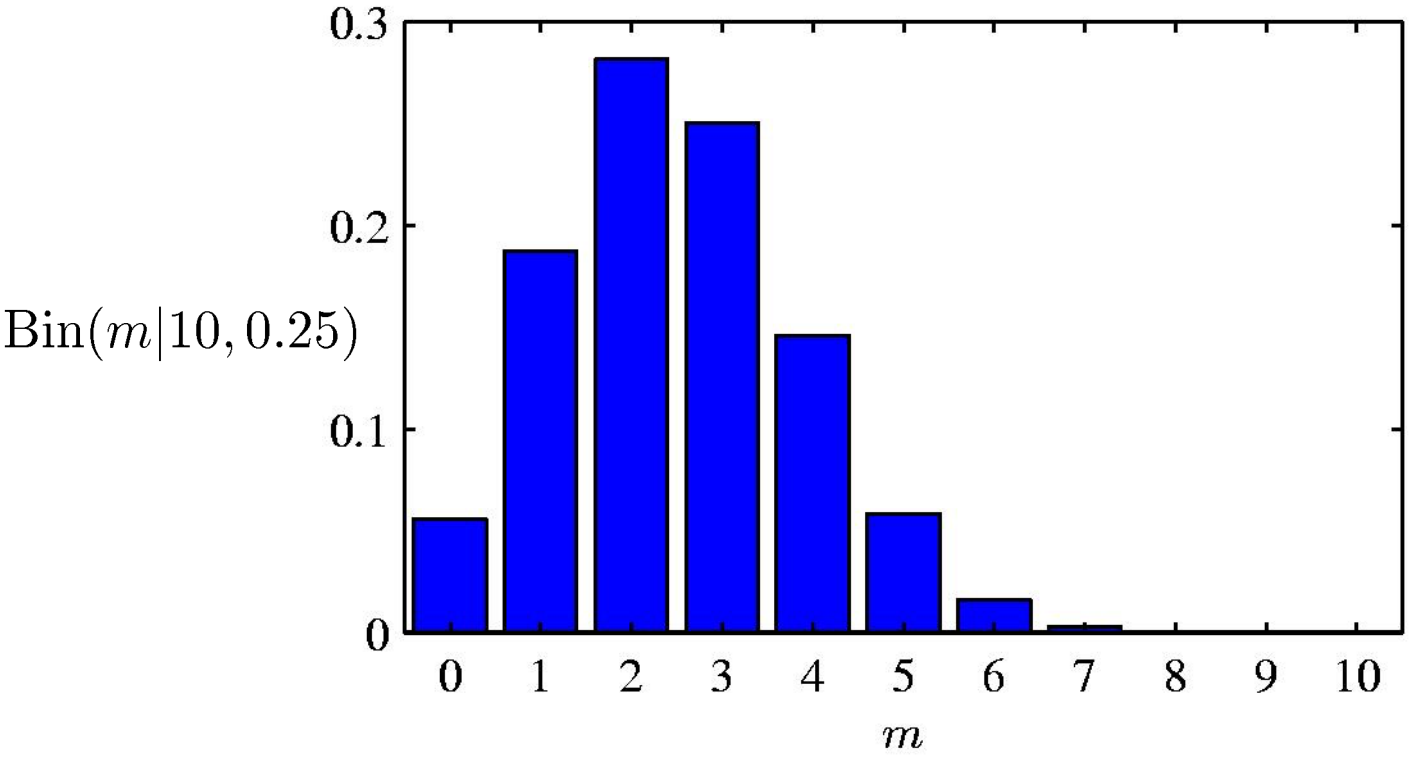
$$\mathbb{E}[X] = \sum_{m=0}^{N} m\binom{N}{m}\mu^m(1-\mu)^{N-m} = N\mu$$

$$\mathrm{var}[X] = \sum_{m=0}^{N} (m-N\mu)^2\binom{N}{m}\mu^m(1-\mu)^{N-m} = N\mu(1-\mu)$$

Verify!

Example

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



Multinomial Distribution

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

• We will use so-called 1-of-K encoding scheme: Only a single element of **x** **is 1, rest is 0.**

• 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:

$$\mathbf{x} = (0, 0, 1, 0, 0, 0)^{\mathrm{T}}$$

x_3

• 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 Distribution

• 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

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

$$\in \mathbb{R}^K$$

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.

k-th entry of $\mathbf{x}_n$

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

- 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 $x_{nk}=1$.

- These are called the sufficient statistics for this distribution.

Maximum Likelihood Estimation

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

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

MLE problem: $\text{maximize} \sum_{k=1}^{K} m_k \ln \mu_k \qquad \text{subject to} \sum_{k=1}^{K} \mu_k = 1$

- 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_{nk}=1.

Multinomial Distribution

• We can construct the joint distribution of the quantities $\{m_1, m_2, \ldots, m_k\}$ given the parameters μ_k and the total number N of observations:

$$\mathrm{Mult}(m_1, m_2, \ldots, m_K | \boldsymbol{\mu}, N) = \binom{N}{m_1 m_2 \ldots m_K} \prod_{k=1}^{K} \mu_k^{m_k}$$

$$\mathbb{E}[m_k] = N\mu_k$$

$$\mathrm{var}[m_k] = N\mu_k(1-\mu_k)$$

$$\mathrm{cov}[m_j m_k] = -N\mu_j\mu_k$$

Verify!

• The normalization coefficient is the number of ways of partitioning N objects into K groups of size $m_1, m_2, \ldots, m_K$.

$$\binom{N}{m_1 m_2 \ldots m_K} = \frac{N!}{m_1! m_2! \ldots m_K!}$$

• Note that $\sum_k m_k = N.$

The Exponential Family

• The exponential families form a basis to many machine learning algorithms, from regression/classification to graphical models.

• The exponential family of distributions over x is defined to be a set of distributions of the form:

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta})\exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\}$$

where

- η is the vector of natural parameters
- u(x) is the vector of sufficient statistics

• The function g(η) can be interpreted as the coefficient that ensures that the distribution p(x|η) is normalized:

$$g(\boldsymbol{\eta})\int h(\mathbf{x})\exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\}\,\mathrm{d}\mathbf{x} = 1$$

Often referred to as partition function.

Bernoulli Distribution

- The Bernoulli distribution is a member of the exponential family:

$$\begin{aligned} p(x|\mu) &= \text{Bern}(x|\mu) = \mu^x(1-\mu)^{1-x} \\ &= \exp\{x\ln\mu + (1-x)\ln(1-\mu)\} \\ &= (1-\mu)\exp\left\{\ln\left(\frac{\mu}{1-\mu}\right)x\right\} \end{aligned}$$

- Comparing with the general form of the exponential family:

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta})\exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\}$$

we see that

$$\eta = \ln\left(\frac{\mu}{1-\mu}\right) \qquad \text{and so} \qquad \mu = \underbrace{\sigma(\eta) = \frac{1}{1+\exp(-\eta)}}_{\text{Sigmoid}}.$$

Bernoulli Distribution

- The Bernoulli distribution is a member of the exponential family:

$$\begin{aligned} p(x|\mu) &= \text{Bern}(x|\mu) = \mu^x(1-\mu)^{1-x} \\ &= \exp\{x\ln\mu + (1-x)\ln(1-\mu)\} \\ &= (1-\mu)\exp\left\{\ln\left(\frac{\mu}{1-\mu}\right)x\right\} \end{aligned}$$

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta})\exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\}$$

- The Bernoulli distribution can therefore be written as:

$$p(x|\eta) = \sigma(-\eta)\exp(\eta x)$$

where

$$\begin{aligned} u(x) &= x \\ h(x) &= 1 \\ g(\eta) &= 1-\sigma(\eta) = \sigma(-\eta). \end{aligned}$$

Multinomial Distribution

- The Multinomial distribution is a member of the exponential family:

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{M} \mu_k^{x_k} = \exp\left\{\sum_{k=1}^{M} x_k \ln \mu_k\right\} = h(\mathbf{x}) g(\boldsymbol{\eta}) \exp\left(\boldsymbol{\eta}^{\mathrm{T}} \mathbf{u}(\mathbf{x})\right)$$

where $\mathbf{x} = (x_1, \ldots, x_M)^{\mathrm{T}}$ $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_M)^{\mathrm{T}}$

and

$$\begin{aligned} \eta_k &= \ln \mu_k \\ \mathbf{u}(\mathbf{x}) &= \mathbf{x} \\ h(\mathbf{x}) &= 1 \\ g(\boldsymbol{\eta}) &= 1. \end{aligned}$$

NOTE: The parameters are constrained to satisfy

$$\sum_{k=1}^{M} \mu_k = 1.$$

This also constrains the natural parameter.

- In some cases it will be convenient to remove the constraint by expressing the distribution over the M-1 parameters.

Multinomial Distribution

- The Multinomial distribution is a member of the exponential family:

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{M} \mu_k^{x_k} = \exp\left\{\sum_{k=1}^{M} x_k \ln \mu_k\right\} = h(\mathbf{x}) g(\boldsymbol{\eta}) \exp\left(\boldsymbol{\eta}^{\mathrm{T}} \mathbf{u}(\mathbf{x})\right)$$

- Let $\mu_M = 1 - \sum_{k=1}^{M-1} \mu_k$ and $x_M = 1 - \sum_{k=1}^{M-1} x_k$
- This leads to:

$$\eta_k = \ln\left(\frac{\mu_k}{1 - \sum_{j=1}^{M-1} \mu_j}\right) \quad \text{and} \quad \mu_k = \underbrace{\frac{\exp(\eta_k)}{1 + \sum_{j=1}^{M-1} \exp(\eta_j)}}_{\text{Softmax function}}.$$

- Here the parameters η_k aren't constrained.
- Note that:

$$0 \leqslant \mu_k \leqslant 1 \quad \text{and} \quad \sum_{k=1}^{M-1} \mu_k \leqslant 1.$$

Multinomial Distribution

- The Multinomial distribution is a member of the exponential family:

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{M} \mu_k^{x_k} = \exp\left\{\sum_{k=1}^{M} x_k \ln \mu_k\right\} = h(\mathbf{x}) g(\boldsymbol{\eta}) \exp\left(\boldsymbol{\eta}^{\mathrm{T}} \mathbf{u}(\mathbf{x})\right)$$

- The Multinomial distribution can therefore be written as:

$$p(\mathbf{x}|\boldsymbol{\mu}) = h(\mathbf{x}) g(\boldsymbol{\eta}) \exp\left(\boldsymbol{\eta}^{\mathrm{T}} \mathbf{u}(\mathbf{x})\right)$$

where

$$\begin{aligned} \boldsymbol{\eta} &= (\eta_1, \ldots, \eta_{M-1}, 0)^{\mathrm{T}} \\ \mathbf{u}(\mathbf{x}) &= \mathbf{x} \\ h(\mathbf{x}) &= 1 \\ g(\boldsymbol{\eta}) &= \left(1 + \sum_{k=1}^{M-1} \exp(\eta_k)\right)^{-1} . \end{aligned}$$

Gaussian Distribution

- The Gaussian distribution can be written as:

$$
\begin{aligned}
p(x|\mu,\sigma^2) &= \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \\
&= \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}x^2+\frac{\mu}{\sigma^2}x-\frac{1}{2\sigma^2}\mu^2\right\} \\
&= h(x)g(\boldsymbol{\eta})\exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(x)\right\}
\end{aligned}
$$

where

$$
\boldsymbol{\eta} = \begin{pmatrix} \mu/\sigma^2 \\ -1/2\sigma^2 \end{pmatrix} \qquad h(\mathbf{x}) = (2\pi)^{-1/2}
$$

$$
\mathbf{u}(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix} \qquad g(\boldsymbol{\eta}) = (-2\eta_2)^{1/2}\exp\left(\frac{\eta_1^2}{4\eta_2}\right).
$$

Mean of Sufficient Statistics

- Remember the Exponential Family:

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\}$$

- From the definition of the normalizer:

$$g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\} \, \mathrm{d}\mathbf{x} = 1$$

- We can take a derivative w.r.t the natural parameter:

$$\nabla g(\boldsymbol{\eta}) \underbrace{\int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\} \, \mathrm{d}\mathbf{x}}_{1/g(\boldsymbol{\eta})} + g(\boldsymbol{\eta}) \underbrace{\int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\} \mathbf{u}(\mathbf{x}) \, \mathrm{d}\mathbf{x}}_{\mathbb{E}[\mathbf{u}(\mathbf{x})]} = 0$$

- Thus

$$-\nabla \ln g(\boldsymbol{\eta}) = \mathbb{E}[\mathbf{u}(\mathbf{x})]$$

Mean of Sufficient Statistics

- Remember the Exponential Family:

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta})\exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\}$$

- We can take a derivative w.r.t η:

$$\nabla g(\boldsymbol{\eta}) \underbrace{\int h(\mathbf{x})\exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\}\,\mathrm{d}\mathbf{x}}_{1/g(\boldsymbol{\eta})} + g(\boldsymbol{\eta}) \underbrace{\int h(\mathbf{x})\exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\}\mathbf{u}(\mathbf{x})\,\mathrm{d}\mathbf{x}}_{\mathbb{E}[\mathbf{u}(\mathbf{x})]} = 0$$

- Thus

$$-\nabla \ln g(\boldsymbol{\eta}) = \mathbb{E}[\mathbf{u}(\mathbf{x})]$$

- Note that the covariance of u(x) can be expressed in terms of the second derivative of g(η), and similarly for the higher moments.
- Just take another derivative! (exercise)

ML for the Exponential Family

- 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 the natural parameter $\boldsymbol{\eta}$.

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\}$$

$$p(\mathbf{X}|\boldsymbol{\eta}) = \left(\prod_{n=1}^{N} h(\mathbf{x}_n)\right) g(\boldsymbol{\eta})^N \exp\left\{\boldsymbol{\eta}^{\mathrm{T}} \sum_{n=1}^{N} \mathbf{u}(\mathbf{x}_n)\right\}.$$

- Therefore we have

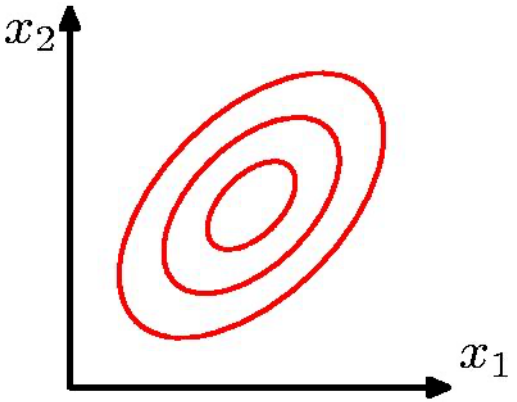
$$-\nabla \ln g(\boldsymbol{\eta}_{\mathrm{ML}}) = \underbrace{\frac{1}{N} \sum_{n=1}^{N} \mathbf{u}(\mathbf{x}_n)}$$

Sufficient Statistic

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.
- $\boldsymbol{\Sigma}$ is a D by D covariance matrix.

and $|\boldsymbol{\Sigma}|$ denotes the determinant of $\boldsymbol{\Sigma}$.

- Note that the covariance is a symmetric **positive definite matrix.**
- Positive definite matrix: $\forall u \in \mathbb{R}^D \quad u^T\boldsymbol{\Sigma}u > 0$
- Positive semidefinite matrix: $\forall u \in \mathbb{R}^D \quad u^T\boldsymbol{\Sigma}u \geq 0$

Means "every" vector in $\mathbb{R}^D$

Moments of the Gaussian Distribution

- The expectation of x under the Gaussian distribution:

$$\begin{aligned}\mathbb{E}[\mathbf{x}] &= \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\int \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}\mathbf{x}\,\mathrm{d}\mathbf{x} \\ &= \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\int \exp\left\{-\frac{1}{2}\mathbf{z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{z}\right\}(\mathbf{z}+\boldsymbol{\mu})\,\mathrm{d}\mathbf{z}\end{aligned}$$

Change of variable:

$$\boldsymbol{x} = \boldsymbol{z} + \boldsymbol{\mu}$$

The term in z in the factor **(z + μ)** will vanish by symmetry, or simply by noticing it is the expectation of a centered Gaussian.

$$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}$$

Moments of the Gaussian Distribution

- The second order moments of the Gaussian distribution:

$$\mathbb{E}[\mathbf{x}\mathbf{x}^{\mathrm{T}}] = \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\Sigma}$$

- The covariance is given by:

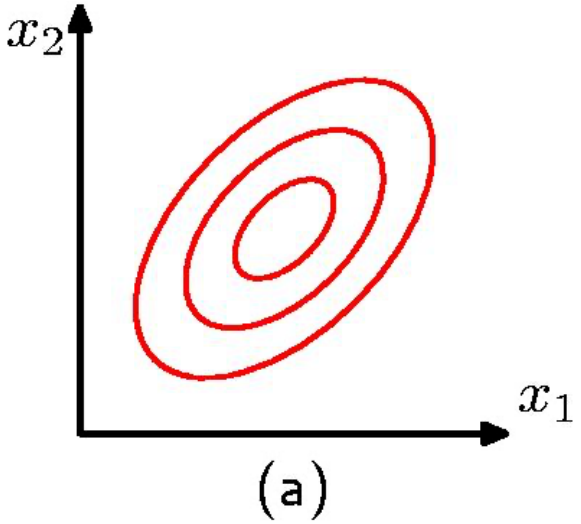
$$\mathrm{cov}[\mathbf{x}] = \mathbb{E}\left[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^{\mathrm{T}}\right] = \boldsymbol{\Sigma}$$

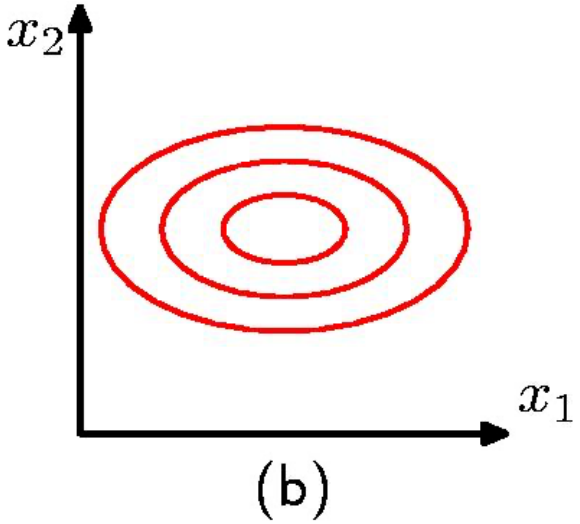
$$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}$$

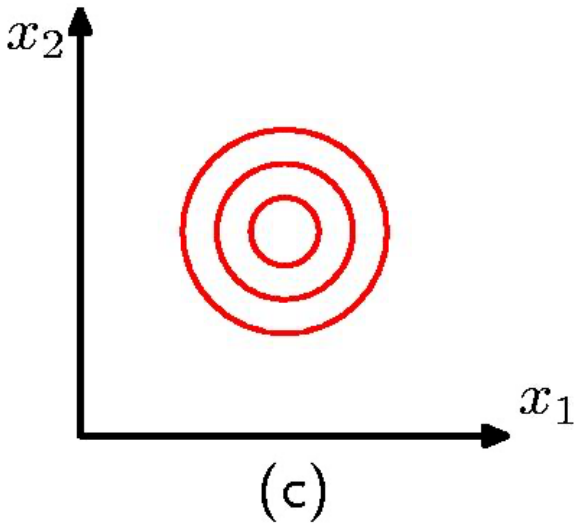
$$\mathbb{E}[\|\mathbf{x}\|^2] = \sum_i \mathbb{E}[\mathbf{x}_i^2] = \mathrm{Tr}(\mathbb{E}[\mathbf{x}\mathbf{x}^T]) = \|\mu\|^2 + \mathrm{Tr}(\boldsymbol{\Sigma})$$

Moments of the Gaussian Distribution

- Contours of constant probability density:







Covariance matrix is of general form.

$$\begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

Diagonal, axis-aligned covariance matrix.

$$\begin{bmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{bmatrix}$$

Spherical (proportional to identity) covariance matrix.

$$\begin{bmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{11} \end{bmatrix}$$

Geometry of the Gaussian Distribution

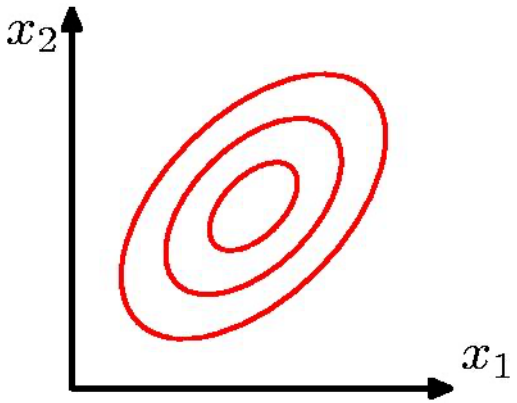
• For a D-dimensional vector x, the Gaussian distribution takes form:

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

• Let us analyze the functional dependence of the Gaussian on x through the quadratic form:

$$\Delta^2 = (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$$

• Here Δ is known as Mahalanobis distance between $\mathbf{x}$ and μ



• The Gaussian distribution will be constant on surfaces in x-space for which Δ is constant.

Maximum Likelihood Estimation

- Suppose we observed i.i.d data $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$.
- We can write the log-likelihood, which is a function of μ and Σ:

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

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

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

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

$$\mathbb{E}[\boldsymbol{\mu}_{\mathrm{ML}}] = \boldsymbol{\mu}$$

$$\mathbb{E}[\boldsymbol{\Sigma}_{\mathrm{ML}}] = \frac{N-1}{N}\boldsymbol{\Sigma}.$$

Biased estimate

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

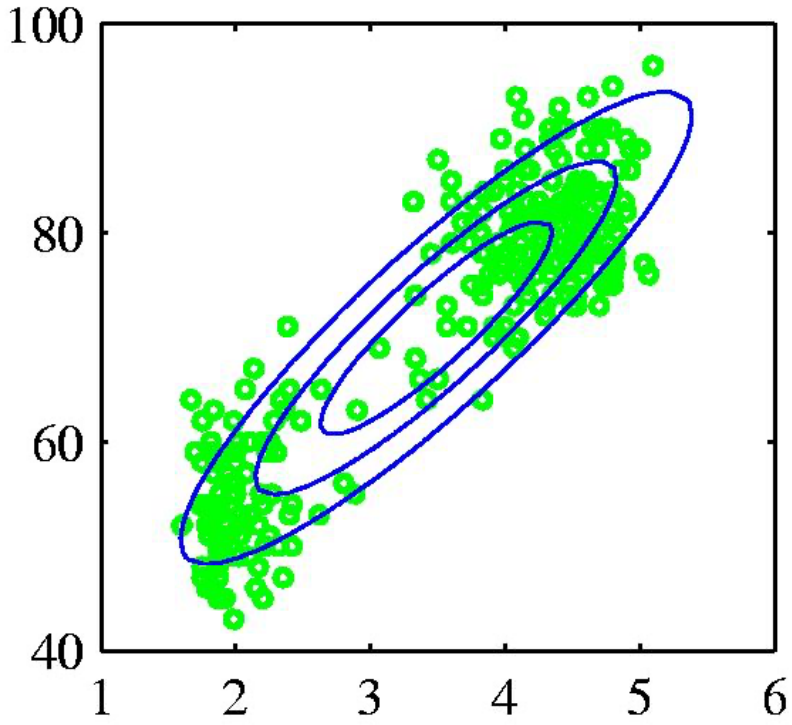
• We can correct the bias by defining a different estimator:

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

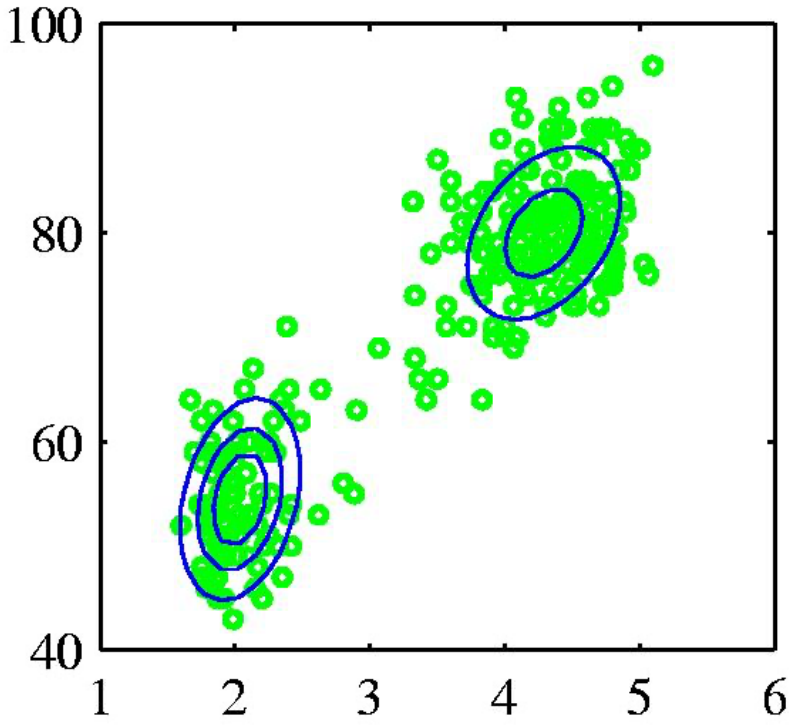
Mixture of Gaussians

- 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

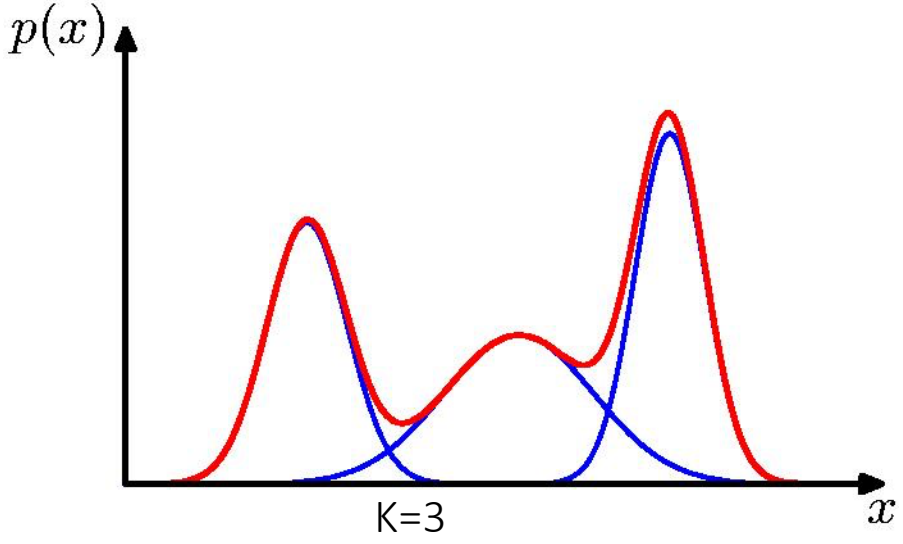
• We can combine simple models into a complex model by defining a superposition of K Gaussian densities of the form:

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

Component

Mixing coefficient

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

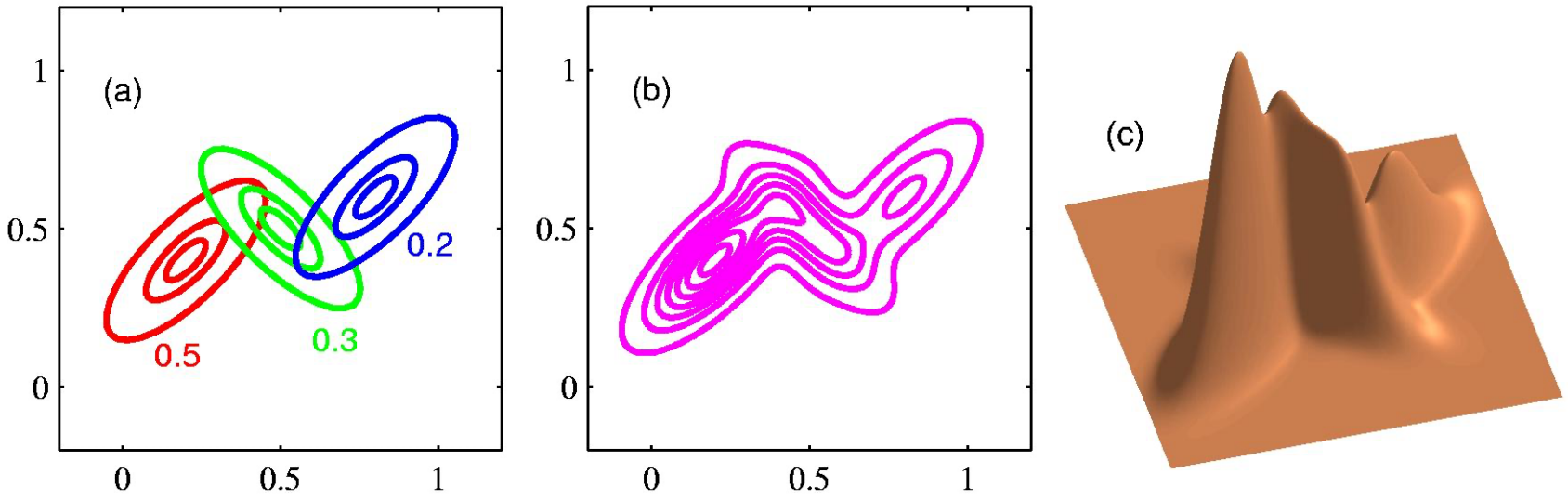


• Note that each Gaussian component has its own mean and covariance. The parameters π_k are called mixing coefficients.

• 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).

Maximum Likelihood Estimation

• Given a dataset D, we can determine model parameters by maximizing the log-likelihood function:

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

Log of a sum: no closed form solution

• Solution: use standard, iterative, numeric optimization methods or the Expectation Maximization algorithm, which we will cover very soon.

Beta Distribution

• We can define a distribution over $\mu \in [0, 1]$ (e.g. it can be used a prior over the parameter μ of the Bernoulli distribution).

$$\begin{aligned} \text{Beta}(\mu|a, b) &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \\ \mathbb{E}[\mu] &= \frac{a}{a+b} \\ \text{var}[\mu] &= \frac{ab}{(a+b)^2(a+b+1)} \end{aligned}$$

where the gamma function is defined as:

$$\Gamma(x) \equiv \int_0^\infty u^{x-1}e^{-u}du.$$

and ensures that the Beta distribution is normalized.

Beta Distribution