CS 412 Introduction to Machine Learning

Probability Theory (1)

Instructor: Wei Tang

Department of Computer Science
University of Illinois at Chicago
Chicago IL 60607

https://tangw.people.uic.edu
tangw@uic.edu

Slides credit: Sargur N. Srihari

Why probability theory?

1. Non-probabilistic models
   – Directly assign $\boldsymbol{x}$ to a specific class
      • E.g., K nearest neighbor
2. Probabilistic Models
   - Discriminative approach
      • Model $p(C_k|\boldsymbol{x})$ in *inference* stage (direct)
      • Use it to make *optimal* decisions
      • E.g., Logistic Regression
   - Generative approach
      • Model class-conditional density $p(\boldsymbol{x}|C_k)$
      • Together with $p(C_k)$ use Bayes rule to compute posterior
      • E.g., Naive Bayes classifier

Probability Theory in Machine Learning

- Probability is key concept is dealing with uncertainty
  - Arises due to finite size of data sets and noise on measurements
- Probability Theory
  - Framework for quantification and manipulation of uncertainty
  - One of the central foundations of machine learning

Random Variable (R.V.)

- Takes values subject to chance
  - E.g., X is the result of coin toss with values $Head$ and $Tail$ which are non - numeric
    - X can be denoted by a r.v. x which has values of 1 and 0
  - Each value of x has an associated probability
- Probability Distribution
  - Mapping from values of a random variable to probabilities
  - p(x=0) = 0.5; p(x=1) = 0.5

Probability with Two Variables

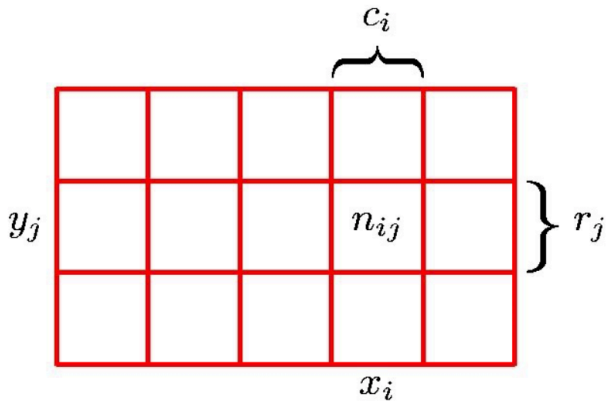
- Key concepts:
  - conditional & joint probabilities of variables
- Random Variables: S and C
  - Shape S, Color C
    - S has two values circle or square
    - C has values red or blue

Probabilities of Interest

- Marginal Probability
  - what is the probability of a square? $P(S=square)$
- Conditional Probability
  - Given that we have a square what is the probability that it is red? $P(C=red|S=square)$
- Joint Probability
  - What is the probability of square AND blue object? $P(S=square,C=blue)$

Sum Rule of Probability Theory

- Consider two random variables
- X can take on values x_i, i=1,, M
- Y can take on values y_i, i=1,..L
- N trials sampling both X and Y
- No of trials with X=x_i and Y=y_j is n_{ij}



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

- Marginal Probability $p(X = x_i) = \frac{c_i}{N}$

$$\text{Since } c_i = \sum_j n_{ij}, \quad \boxed{p(X = x_i) = \sum_{j=1}^{L} p(X = x_i, Y = y_j)}$$

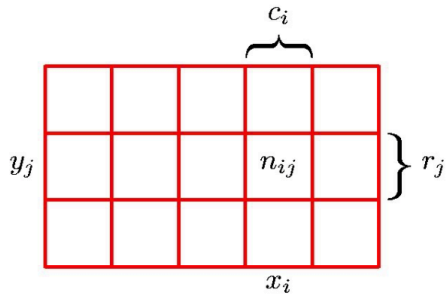
Product Rule of Probability Theory

- Consider only those instances for which $X=x_i$
- Then fraction of those instances for which $Y=y_j$ is written as $p(Y=y_j|X=x_i)$
- Called conditional probability
- Relationship between joint and conditional probability:

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

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

$$= \boxed{p(Y = y_j \mid X = x_i)p(X = x_i)}$$



Bayes Theorem

- From the product rule together with the symmetry property $p(X,Y)=p(Y,X)$ we get

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

$$Posterior \propto Likelihood \times Prior$$

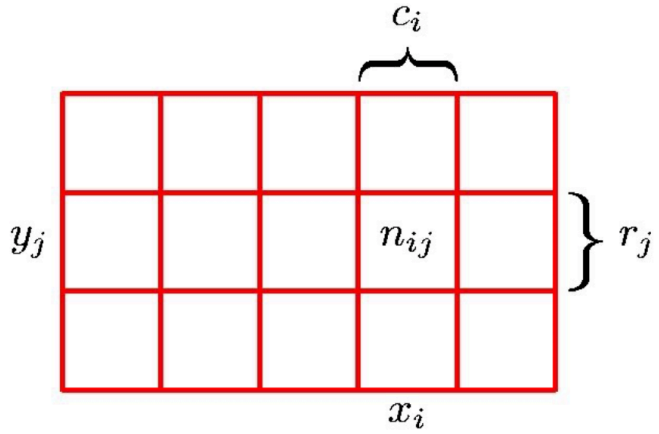
- Which is called Bayes' theorem
- Using the sum rule the denominator is expressed as

$$p(X) = \sum_{Y} p(X \mid Y)p(Y)$$

Normalization constant to ensure sum of conditional probability on LHS sums to 1 over all values of Y

Rules of Probability

- Given random variables X and Y
- Sum Rule gives Marginal Probability



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

Type equation here.

- Product Rule: joint probability in terms of conditional and marginal

$$p(X,Y) = \frac{n_{ij}}{N} = p(Y \mid X)p(X) = \frac{n_{ij}}{c_i} \times \frac{c_i}{N}$$

- Combining we get Bayes Rule

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

where

$$p(X) = \sum_{Y} p(X \mid Y)p(Y)$$

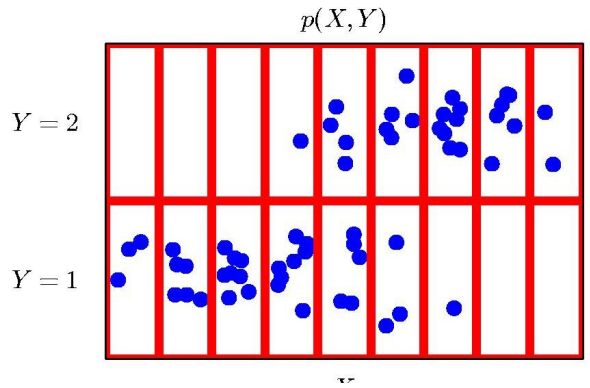
Viewed as

$$Posterior \propto Likelihood \times Prior$$

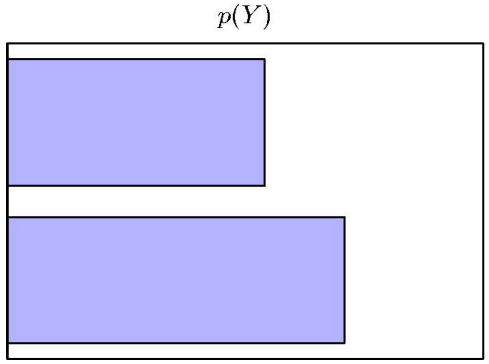
Ex: Joint Distribution over two Variables

X takes nine possible values, Y takes two values

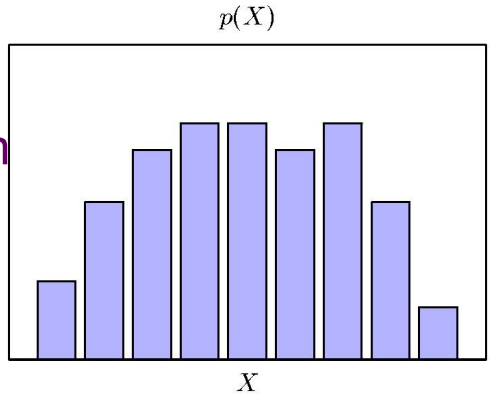
$N = 60$ data points



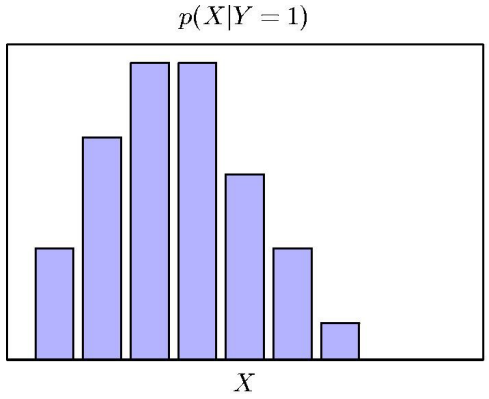
Histogram of Y (Fraction of data points having each value of Y)



Histogram of X



Histogram of X given Y=1



Fractions would equal the probability as N -> ∞

Independent and Dependent Variables

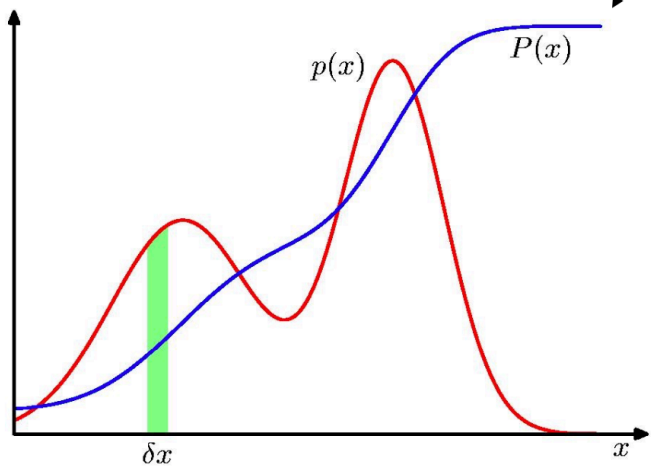
- Independent variables
  - If $p(X,Y)=p(X)p(Y)$ then X and Y are independent
  - Why?
  - From product rule: $$p(Y \mid X) = \frac{p(X,Y)}{p(X)} = p(Y)$$
- Dependent variables
  - Variables are not independent

Probability Density Function (pdf)

- Continuous Variables
- If probability that x falls in interval $(x, x + \delta x)$ is given by $p(x)\delta x$ for $\delta x \rightarrow 0$ then $p(x)$ is a pdf of x
- Probability x lies in interval (a,b) is

$$p(x \in (a,b)) = \int_a^b p(x)\,dx$$

Cumulative Distribution Function



Probability that x lies in Interval $(-\infty,z)$ is

$$P(z) = \int_{-\infty}^{z} p(x)\,dx$$

Several Variables

- If there are several continuous variables $x_1,\ldots,x_D$ denoted by vector $\mathbf{x}$ then we can define a joint probability density $p(\mathbf{x})=p(x_1,..,x_D)$
- Multivariate probability density must satisfy

$$p(\mathbf{x}) \geq 0$$

$$\int_{-\infty}^{\infty} p(\mathbf{x})\, d\mathbf{x} = 1$$

Sum, Product, Bayes for Continuous

- Rules apply for continuous, or combinations of discrete and continuous variables

$$p(x) = \int p(x, y)\, dy$$

$$p(x, y) = p(y \mid x) p(x)$$

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

Expectation

- Expectation is *average* value of some function $f(x)$ under the probability distribution $p(x)$ denoted $E[f]$
- For a discrete distribution

$$E[f] = \sum_x p(x)\, f(x)$$

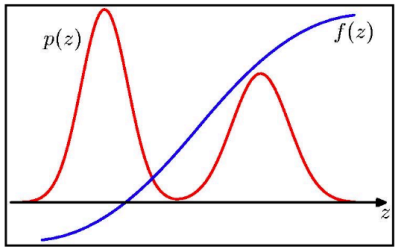
- For a continuous distribution

$$E[f] = \int p(x) f(x)\, dx$$

- If there are N points drawn from a pdf, then expectation can be approximated as

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

Examples of $f(x)$ of use in ML:



$f(x)=x$; $E[f]$ is mean

$f(x)=\ln p(x)$; $E[f]$ is entropy

$f(x)=-\ln[q(x)/p(x)]$; K-L divergence

This approximation is extremely important when we use sampling to determine expected value

Variance

- Measures how much variability there is in $f(x)$ around its mean value $E[f(x)]$
- Variance of $f(x)$ is denoted as

$$\text{var}[f] = E[(f(x) - E[f(x)])^2]$$

- *Expanding the square*

$$\text{var}[f] = E[(f(x)^2] - E[f(x)]^2$$

- Variance of the variable x itself

$$\text{var}[x] = E[x^2] - E[x]^2$$

Covariance

- For two random variables x and y their covariance is
- $$\text{cov}[x,y] = E_{x,y}[\{x-E[x]\}\{y-E[y]\}]$$
  $$= E_{x,y}[xy] - E[x]E[y]$$
  - Expresses how x and y vary together
- If x and y are independent then their covariance vanishes
- If x and y are two vectors of random variables covariance is a matrix
- If we consider covariance of components of vector x with each other then we denote it as $\text{cov}[x] = \text{cov}[x,x]$

Bayesian Probabilities

- Classical or Frequentist view of Probabilities
  - Probability is frequency of random, repeatable event
  - Frequency of a tossed coin coming up heads is 1/2
- Bayesian View
  - Probability is a quantification of uncertainty
  - Degree of belief in propositions that do not involve random variables
  - Examples of uncertain events as probabilities:
    - Whether Arctic Sea ice cap will disappear
    - Whether moon was once in its own orbit around the sun
    - Whether Thomas Jefferson had a child by one of his slaves
    - Whether a signature on a check is genuine

Bayesian Approach

- Quantify uncertainty around choice of parameters $\mathbf{w}$
  - E.g., $\mathbf{w}$ is vector of parameters in curve fitting

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + .. + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

- Uncertainty before observing data expressed by $p(\mathbf{w})$
- Given observed data $D = \{ t_1, . . t_N \}$
  - Uncertainty in $\mathbf{w}$ after observing D, by Bayes rule:

$$p(\mathbf{w} \mid D) = \frac{p(D \mid \mathbf{w}) p(\mathbf{w})}{p(D)}$$

  - Quantity $p(D/\mathbf{w})$ is evaluated for observed data
    - It can be viewed as function of $\mathbf{w}$
    - It represents how probable the data set is for different parameters $\mathbf{w}$
    - It is called the *Likelihood function*
    - Not a probability distribution over $\mathbf{w}$

Bayes theorem in words

- Uncertainty in **w** expressed as

$$p(\mathbf{w} \mid D) = \frac{p(D \mid \mathbf{w})p(\mathbf{w})}{p(D)}$$

- Bayes theorem in words:

  posterior α likelihood $\times$ prior

- Denominator is normalization factor
  - Involves marginalization over **w**

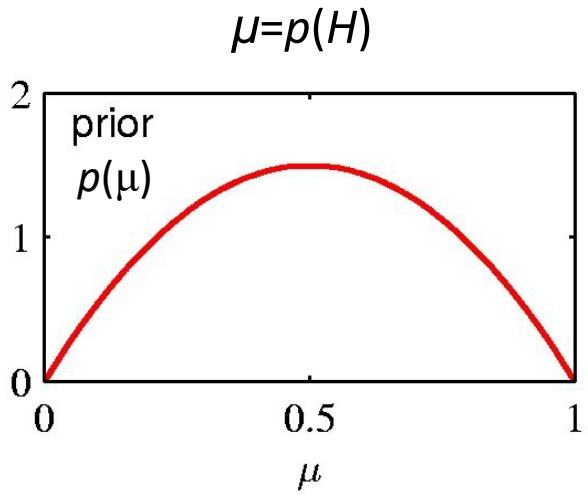
$$p(D) = \int p(D \mid \mathbf{w})p(\mathbf{w})\,d\mathbf{w} \quad \text{by Sum Rule}$$

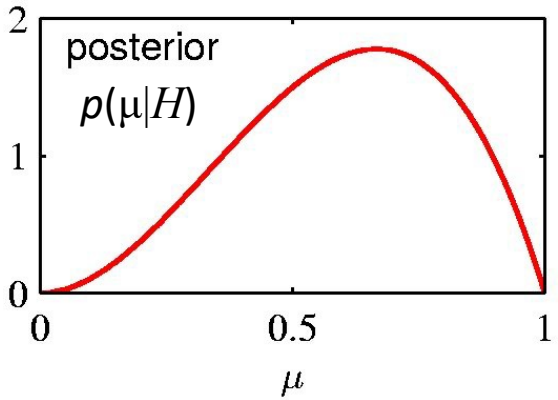
Maximum Likelihood Approach

- In frequentist setting $\mathbf{w}$ is a fixed parameter
  - $\mathbf{w}$ is set to value that maximizes likelihood function $p(D|\mathbf{w})$
  - In ML, negative log of likelihood function is called error function since maximizing likelihood is equivalent to minimizing error

Bayesian: Prior and Posterior

- Inclusion of prior knowledge arises naturally
- Coin Toss Example
  - Fair looking coin is tossed three times and lands Head each time
  - Classical MLE of the probability of landing heads is 1 implying all future tosses will land *Heads*
  - Bayesian approach with reasonable prior will lead to less extreme conclusion





Bernoulli distribution

The discrete probability distribution of a random variable which takes the value 1 with probability p and the value 0 with probability q=1-p.

$$P(x = 1) = p$$
$$P(x = 0) = 1 - p$$

Multinoulli distribution

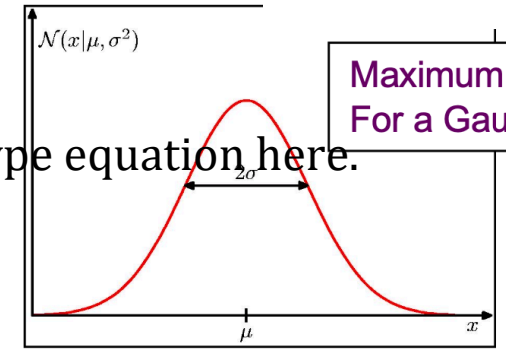
- Also called generalized Bernoulli distribution or categorical distribution
- K categories
- $P(x = k) = p_k$
- $p_1 + \cdots + p_K$=1

The Gaussian Distribution

- For single real-valued variable ***x***

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

- It has two parameters:
  - Mean μ, variance σ^2,
  - *Standard deviation* σ
    - *Precision* $\beta = 1/\sigma^2$



Type equation here.

Maximum of a distribution is its mode
For a Gaussian, mode coincides with mean

- Can find expectations of functions of ***x*** under Gaussian

$$E[x] = \int_{-\infty}^{\infty} N(x \mid \mu, \sigma^2) x dx = \mu$$

$$E[x^2] = \int_{-\infty}^{\infty} N(x \mid \mu, \sigma^2) x^2 dx = \mu^2 + \sigma^2$$

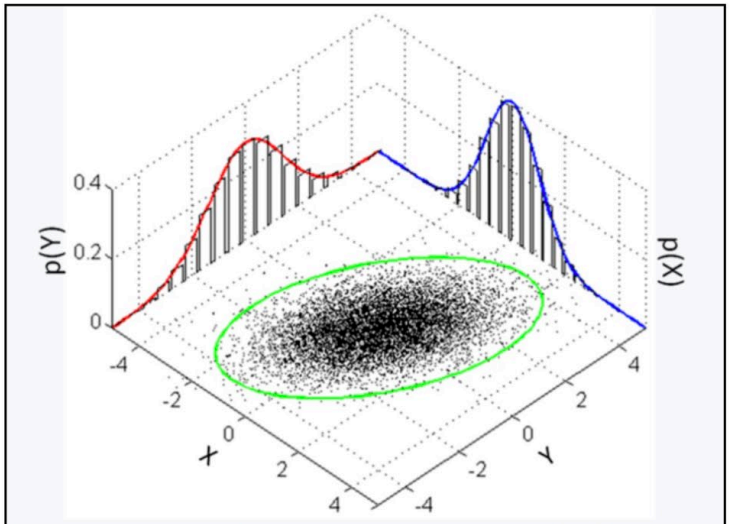
$$\mathrm{var}[x] = E[x^2] - E[x]^2 = \sigma^2$$

Multivariate Gaussian Distribution

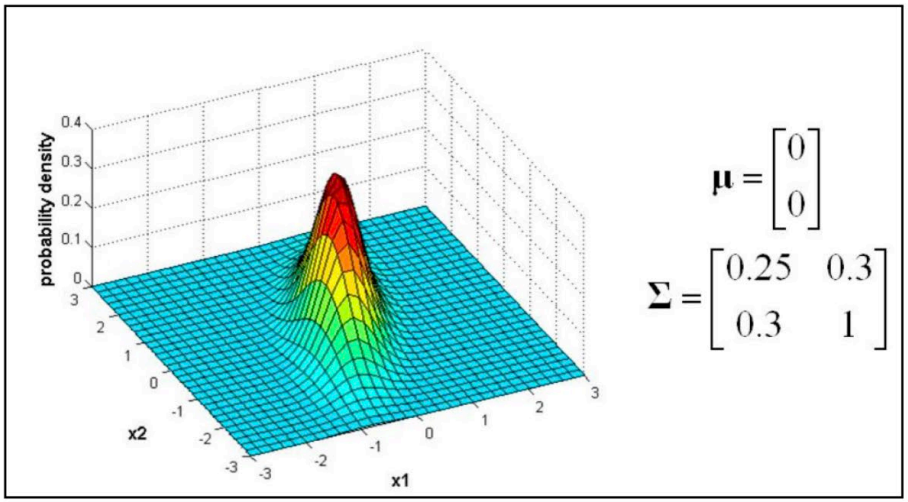
- For a real-valued random vector x

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

- It has parameters:
  - Mean $\boldsymbol{\mu}$, a D-dimensional vector
  - Covariance matrix Σ
    - Which is a $D \times D$ matrix



Many sample points from a multivariate normal distribution with $\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 3/5 \\ 3/5 & 2 \end{bmatrix}$, shown along with the 3-sigma ellipse, the two marginal distributions, and the two 1-d histograms.



Central limit theorem

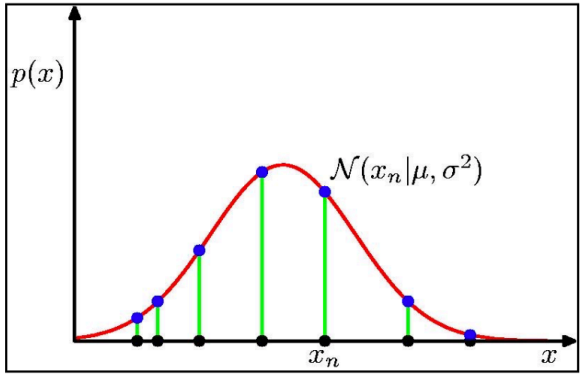
Lindeberg–Lévy CLT. Suppose $\{X_1, \ldots, X_n\}$ is a sequence of i.i.d. random variables with $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2 < \infty$. Then as n approaches infinity, the random variables $\sqrt{n}(\bar{X}_n - \mu)$ converge in distribution to a normal $\mathcal{N}(0, \sigma^2)$:[4]

$$\sqrt{n}\left(\bar{X}_n - \mu\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right).$$

Likelihood Function for Gaussian

- Given *N* scalar observations $\mathbf{x}=[x_1,..\ x_n]^T$
  - Which are independent and identically distributed
- Probability of data set is given by likelihood function

$$p(\mathrm{x} \mid \mu, \sigma^2) = \prod_{n=1}^{N} N(x_n \mid \mu, \sigma^2)$$



Data: black points
Likelihood= product of blue values
Pick mean and variance to maximize this product

- Log-likelihood function is

$$\ln p(\mathrm{x} \mid \mu, \sigma^2) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)$$

- Maximum likelihood solutions are given by

$$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n$$ which is the sample mean

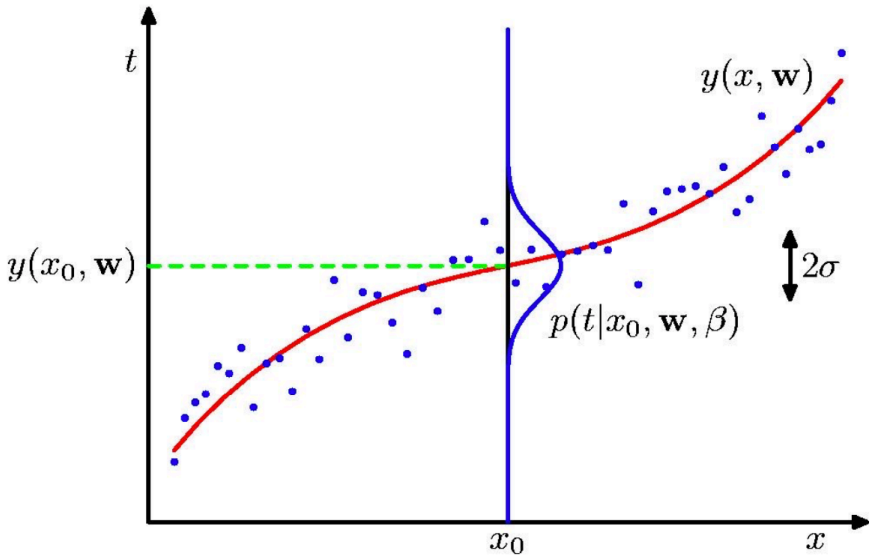
$$\sigma^2_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{ML})^2$$ which is the sample variance

Curve Fitting Probabilistically

- Goal is to predict for target variable *t* given a new value of the input variable *x*
  - Given N input values $\mathbf{x}=(x_1,..x_N)^T$ and corresponding target values $\mathbf{t}=(t_1,..,t_N)^T$
  - Assume given value of *x*, value of *t* has a Gaussian distribution with mean equal to $y(x,\mathbf{w})$ of polynomial curve

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

$$y(x,\mathbf{w}) = w_0 + w_1 x + w_2 x^2 + .. + w_M x^M = \sum_{j=0}^{M} w_j x^j$$



Gaussian conditional distribution for *t* given *x*.
Mean is given by polynomial function $y(x,\mathbf{w})$
Precision given by β

Curve Fitting with Maximum Likelihood

- Likelihood Function is $$p(\mathbf{t} \mid \mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} N(t_n \mid y(x_n, \mathbf{w}), \beta^{-1})$$
- Logarithm of the Likelihood function is
$$\ln p(\mathbf{t} \mid \mathbf{x}, \mathbf{w}, \beta) = -\frac{\beta}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi)$$
- To find maximum likelihood solution for polynomial coefficients $\mathbf{w}_{ML}$
  - Maximize w.r.t $\mathbf{w}$
  - Can omit last two terms -- don't depend on $\mathbf{w}$
  - Can replace $\beta/2$ with ½ (since it is constant wrt $\mathbf{w}$)
  - Minimize negative log-likelihood
  - Identical to sum-of-squares error function

Precision parameter with MLE

- Maximum likelihood can also be used to determine β of Gaussian conditional distribution
- Maximizing likelihood wrt β gives

$$\frac{1}{\beta_{\mathrm{ML}}} = \frac{1}{N}\sum_{n=1}^{N}\left\{y(x_n, \mathbf{w}_{\mathrm{ML}}) - t_n\right\}^2$$

- First determine parameter vector $\mathbf{w}_{\mathrm{ML}}$ governing the mean and subsequently use this to find precision β_{ML}

Predictive Distribution

- Knowing parameters $\mathbf{w}$ and β
- Predictions for new values of x can be made using

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

- Instead of a point estimate we are now giving a probability distribution over t

A More Bayesian Treatment

- Introducing a prior distribution over polynomial coefficients **w**

$$p(\mathbf{w} \mid \alpha) = N(\mathbf{w} \mid 0, \alpha^{-1} I) = \left(\frac{\alpha}{2\pi}\right)^{(M+1)/2} \exp\left\{-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}\right\}$$

  - where α is the precision of the distribution
  - M+1 is total no. of parameters for an M^{th} order polynomial
  - α are Model parameters also called *hyperparameter*
    - they control distribution of model parameters

Posterior Distribution

- Using Bayes theorem, posterior distribution for $\mathbf{w}$ is proportional to product of prior distribution and likelihood function

$$p(\mathbf{w}|\mathbf{x},\mathbf{t},\alpha,\beta) \quad \propto \quad p(\mathbf{t}|\mathbf{x},\mathbf{w},\beta)p(\mathbf{w}|\alpha)$$

- $\mathbf{w}$ can be determined by finding the most probable value of $\mathbf{w}$ given the data, ie. maximizing posterior distribution
- This is equivalent (by taking logs) to minimizing

$$\frac{\beta}{2}\sum_{n=1}^{N}\left\{y(x_n,\mathbf{w}\ )-t_n\right\}^2+\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}$$

- Same as sum of squared errors function with a regularization parameter given by $\lambda=\alpha/\beta$

Models in Curve Fitting

- In polynomial curve fitting:
  - an optimal order of polynomial gives best generalization
- Order of the polynomial controls
  - the number of free parameters in the model and thereby model complexity
- With regularized least squares λ also controls model complexity