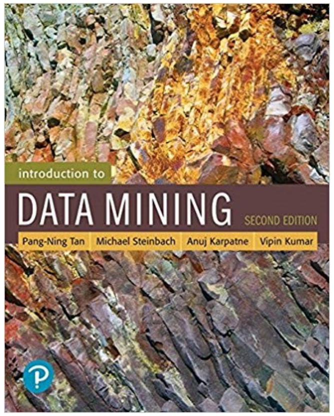


CIS 530—Advanced Data Mining

5- Probability Theory

Thomas W Gyeera, Assistant Professor
Computer and Information Science
University of Massachusetts Dartmouth

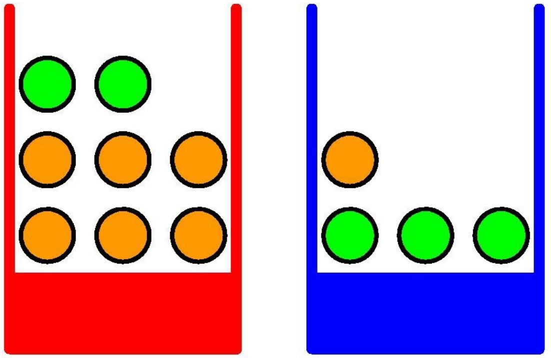
Courtesy to Prof. Sargur N. Srihari

Probabilities of Interest

- Marginal Probability
  - What is the probability of an apple?
- Conditional Probability
  - Given that we have an orange what is the probability that we chose the blue box?
- Joint Probability
  - What is the probability of orange AND blue box?

2 apples
6 oranges

3 apples
1 orange

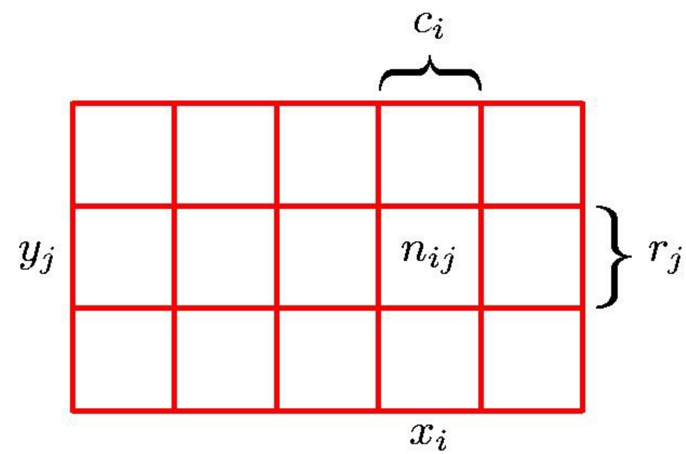


Box is random variable B
(has values r or b)
Fruit is random variable F
(has values o or a)

Let
and

Sum Rule of Probability Theory

- Consider two random variables
  - can take on values
  - can take on values
  - trials sampling both and
  - No. of trials with and is



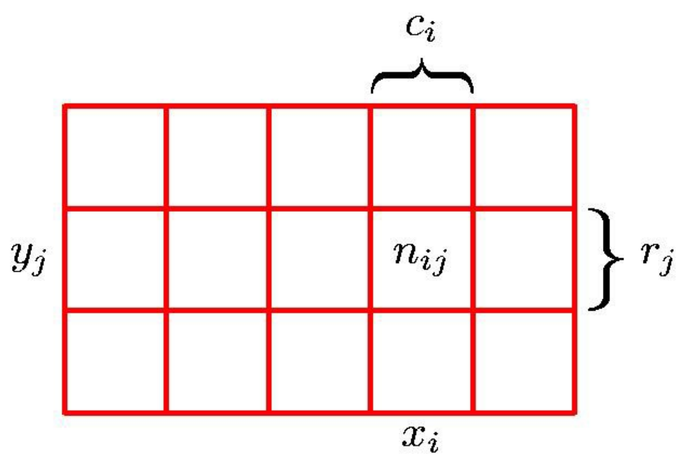
Joint Probability $p(X = x_i, \; Y = y_j) = \dfrac{n_{ij}}{N}$

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

Since $c_i = \sum_j n_{ij}$ $\quad p(X = x_i) = \sum_{j=1}^{3} p(X = x_i, Y = y_j)$

Product Rule of Probability Theory

- Consider only those instances for which
- Then fraction of those instances for which is written as
- Called conditional probability
- Relationship between joint and conditional probability:



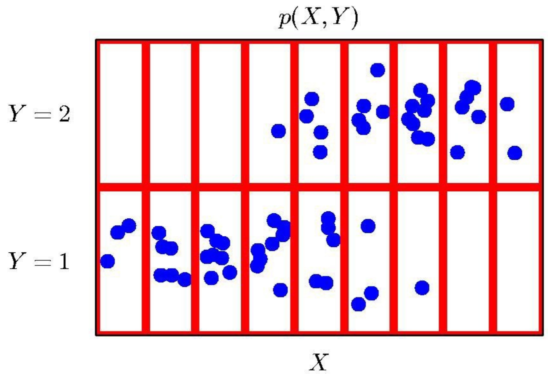
$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$

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

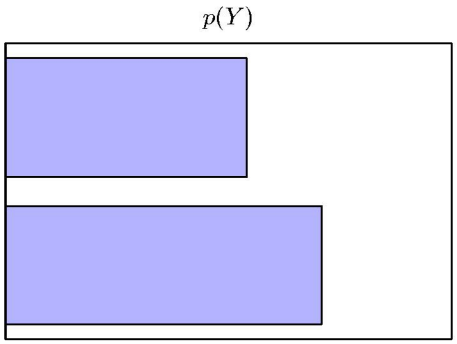
$$= p(Y = y_j | X = x_i)\, p(X = x_i)$$

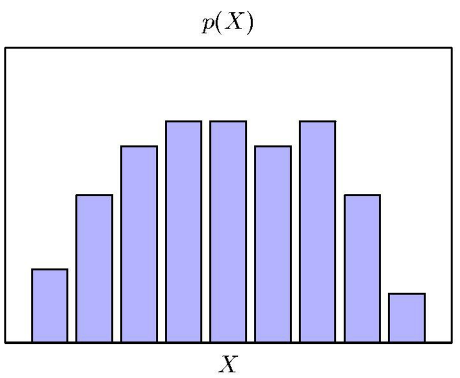
Joint Distribution over Two Variables

N = 60 data points

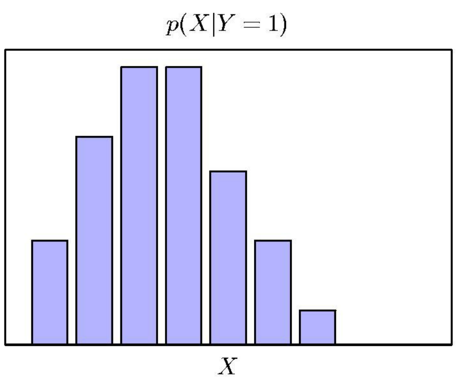


Histogram of Y
(Fraction of
data points
having each
value of Y)





Histogram of X



Histogram
of X given Y=1

Fractions would equal the probability as N ☾ oo

Bayes Theorem

- Think about relation between mid-term score and final grade:
  - Everyone wants to know, if my mid-term is in the range [80-90], what is the probability of getting A?
  - Assume the range of score is represented by random variable , and final grade is by
  - So, we are modeling:
- Below is something we may know from the course records of year 2018:

Bayes Theorem

- From the product rule together with the symmetry property we get

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

- which is called Bayes' theorem
- Sum and product rule for

$$p(X) = \sum_{Y} p(X|Y)\,p(Y)$$

Normalization constant to ensure sum of conditional probability equal to 1 over all values of Y

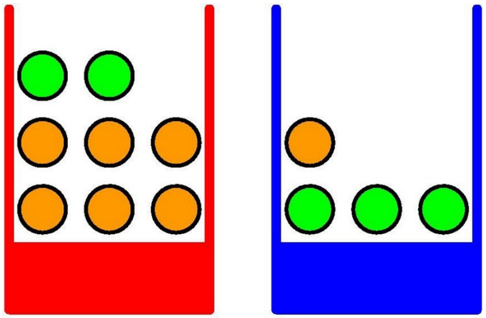
Bayes rule applied to Fruit Problem

- Probability that box is red given that fruit picked is orange

$$p(B = r \mid F = o) = \frac{p(F = o \mid B = r)\, p(B = r)}{p(F = o)}$$

$$= \frac{\frac{3}{4} \times \frac{4}{10}}{\frac{9}{20}} = \boxed{\frac{2}{3} = 0.66}$$

Let
and



The *a posteriori* probability of 0.66 is different from the *a priori* probability of 0.4

- Probability that fruit is orange
  - From sum and product rules

$$p(F = o) = p(F = o, B = r) + p(F = o, B = b)$$

$$= p(F = o \mid B = r)\, p(B = r) + p(F = o \mid B = b)\, p(B = b)$$

$$= 6 \times 4 + 1 \times 6 = \boxed{\frac{9}{20} = 0.45}$$

The *marginal* probability of 0.45 is lower than average probability of 7/12=0.58

Independent Variables

- If then and are said to be independent
- Why?
- From product rule

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

- In fruit example if each box contained same fraction of apples and oranges then

Can you prove this?

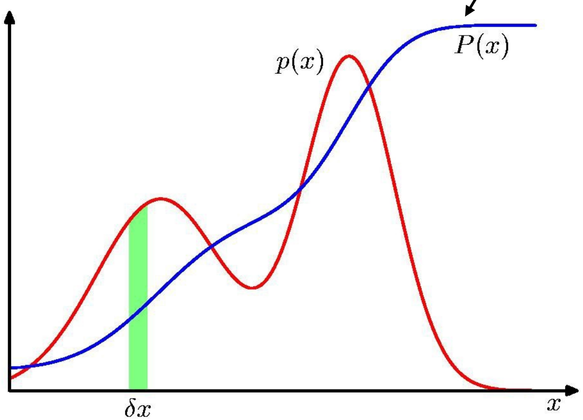
Probability Densities Function (PDF)

- Continuous Variables
  - If probability that falls in interval is given by
    for
  - then is a Probability Density Function (PDF) of
- Probability lies in interval is

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

Cumulative

Distribution Function



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

Probability that x lies in Interval *(-oo,z)* is

Several Variables

- If there are several continuous variables denoted by vector then we can define a joint probability density

$$p(\mathrm{x}) = p(x_1, x_2, \ldots, x_D)$$

- Multivariate probability density must satisfy

$$\int_{-\infty}^{\infty} p(\mathrm{x})\, d\mathrm{x} = 1 \qquad p(\mathrm{x}) \geq 0$$

Sum, Product, Bayes for Continuous

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

$$\textbf{Marginalization}\ p(x) = \int p(x,y)\,dy$$

$$\textbf{Product}\ p(x,y) = p(y|x)\,p(x)$$

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

- Formal justification of sum, product rules for continuous variables requires measure theory

Expectation: an Example

- Assume the final grade is represented by a random variable
- Blow is the marginal probability of in class
- We further assume there is a mapping between final grade and future salary level
- Question: what is the expected salary level of this class?

Expectation of

- Expectation is *average* value of some function

  under the probability distribution

- For a discrete distribution: $E[f] = \sum p(x) f(x)$

- For a continuous distribution: $E[f] = \int p(x) f(x) dx$

- If there are points drawn from a PDF, then

  expectation can be approximated as: $E[f] = \frac{1}{N} \sum f(x)$

Variance of and

- Measures how much variability there is in around its mean value
- Variance of is denoted as

$$var[f] = E[f(x) - E[f(x)]]^2$$

- Expanding the square, we have

$$var[f] = E[f^2(x)] - E[f(x)]^2$$

- Variance of the variable itself

$$var[x] = E[x^2] - E[x]^2$$

Covariance

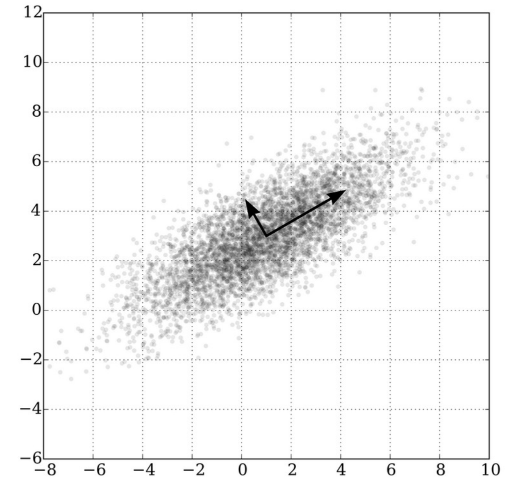
- For two random variables and covariance is defined as

$$cov[x,y] = E_{xy}(x - E[x])(y - E[y])$$

$$= E_{xy}[xy] - E[x]E[y]$$

- Expresses how and vary together
- If and are independent, then their covariance vanishes
- If we consider covariance of components of vector with each other then we denote it as

$$var[x] = cov[x,x]$$



Why??

Covariance Matrix

- If and are two vectors of random variables, then the covariance is a matrix
- Example:
  - assume random variable vectors

$$cov(\mathrm{x},\mathrm{y}) = \begin{pmatrix} cov(x_1,y_1) & cov(x_1,y_2) & cov(x_1,y_3) & cov(x_1,y_4) \\ cov(x_2,y_1) & cov(x_2,y_2) & cov(x_2,y_3) & cov(x_2,y_4) \\ cov(x_3,y_1) & cov(x_3,y_2) & cov(x_3,y_3) & cov(x_3,y_4) \\ cov(x_4,y_1) & cov(x_4,y_2) & cov(x_4,y_3) & cov(x_4,y_4) \end{pmatrix}$$

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 Probabilities

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 Shakespeare's plays were written by Francis Bacon
- 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

Examples of Uncertain Events

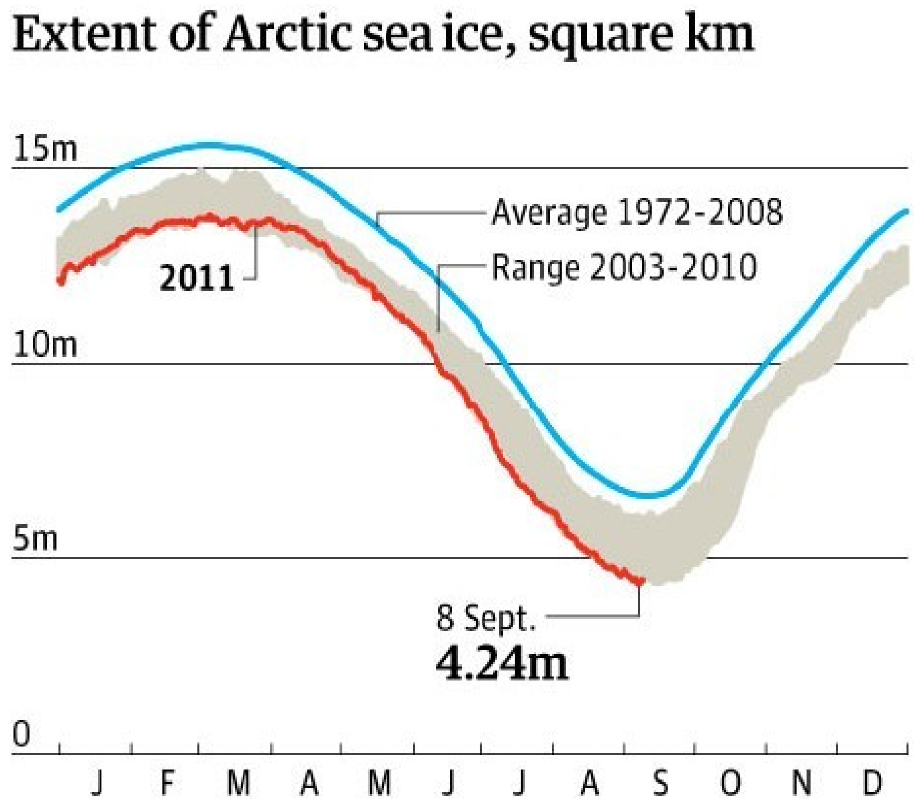
Probability that Mr. M was the murderer of Mrs. M given the evidence

Whether Arctic ice cap will disappear by end of century

- We have some idea of how quickly polar ice is melting
- Revise it on the basis of fresh evidence (satellite observations)
- Assessment will affect actions we take (to reduce greenhouse gases)

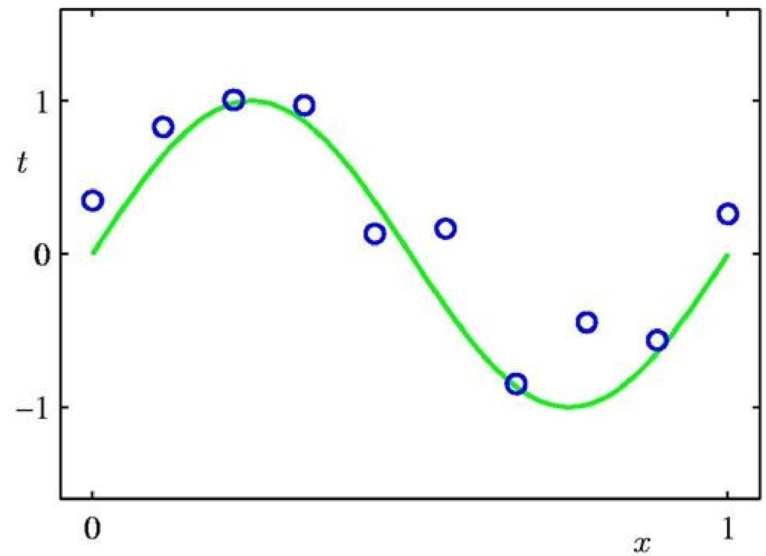
All can be achieved by general Bayesian interpretation

Arctic Ice (2011)



Bayesian Approach

- Quantify uncertainty around choice of parameters
  - E.g., i g



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

- Uncertainty before observing data expressed by

Bayesian Approach

- Given observed data
- Uncertainty in after observing , by Bayes rule:

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

- Quantity can be viewed as a function of
- Represent how probable the data set is for different parameters Called **Likelihood function**
- Not a probability distribution over

Role of Likelihood Function

- Uncertainty in expressed as

$$\boxed{p(\mathbf{w}|D)=\frac{p(D|\mathbf{w})\,p(\mathbf{w})}{p(D)}}$$ where $$\boxed{p(D)=\int p(D|\mathbf{w})\,p(\mathbf{w})\,d\,\mathbf{w}}$$

- Denominator is normalization factor
- Involves marginalization over
- Bayes theorem in words

$$posterior likelihood\ \ prior$$

Role of Likelihood Function

- Likelihood Function plays central role in both Bayesian and frequentist paradigms
  - Frequentist: is a **fixed** parameter determined by an estimator; error bars on estimate are obtained from possible data sets
  - Bayesian: there is a single data set ; uncertainty is expressed as **probability distribution** over

Maximum Likelihood Approach

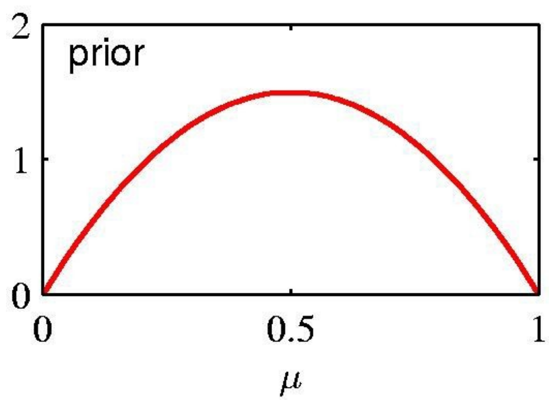
- In frequentist setting, is considered to be a fixed parameter
  - is set to value that maximizes likelihood function
- In ML, negative log of likelihood function is called error function, since maximizing likelihood is equivalent to minimizing error

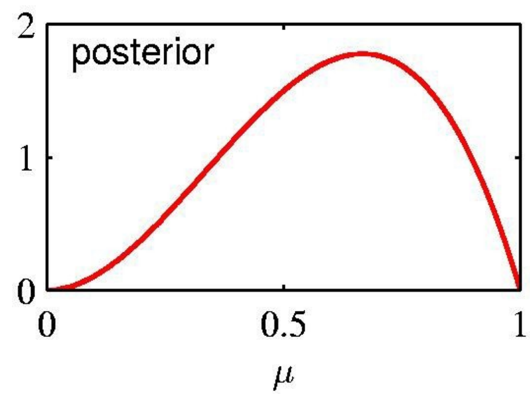
Maximum Likelihood Approach

- Bootstrap approach to creating data sets
  - From data points new data sets are created by drawing points at random with replacement
  - Repeat times to generate data sets
  - Accuracy of parameter estimate can be evaluated by variability of predictions between different bootstrap sets

Bayesian vs. Frequentist Approach

- 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





Practicality of Bayesian Approach

Marginalization over whole parameter space is required to make predictions or compare models

Factors making it practical:

- Sampling Methods such as Markov Chain Monte Carlo methods
- Increased speed and memory of computers

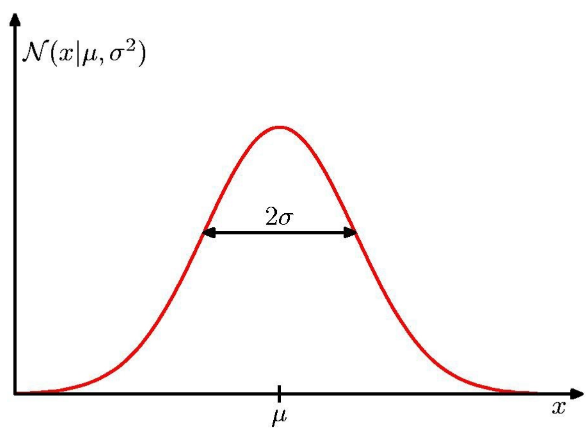
Deterministic approximation schemes such as Variational Bayes and Expectation propagation are alternatives to sampling

The Gaussian Distribution

- For single real-valued variable

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

- Parameters:
  - Mean , variance
  - Standard deviation
  - Precision =

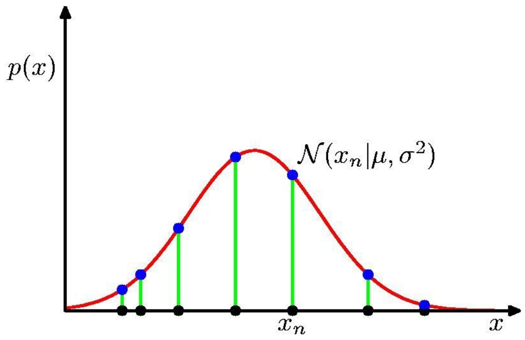


Maximum of a distribution is its mode

For a Gaussian, mode coincides with its mean

Likelihood Function for Gaussian

- Given observations
- Independent and identically distributed (i.i.d.)
- Probability of data set is given by likelihood function



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

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

- Log-likelihood function is

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

Likelihood Function for Gaussian

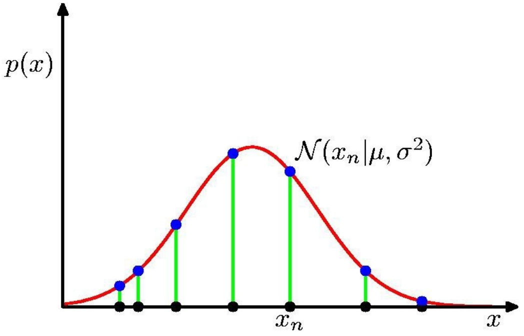
- Log-likelihood function is

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

- Maximum likelihood solutions are given by:

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

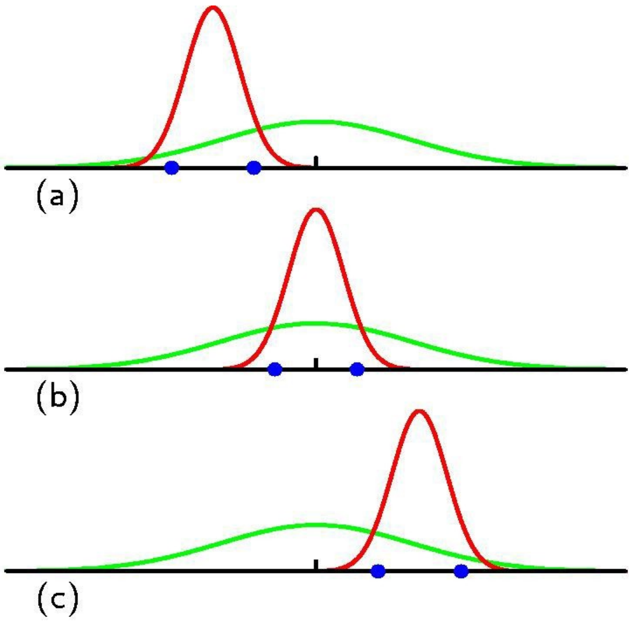
$$\sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2$$



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

Bias in Maximum Likelihood

- Maximum likelihood systematically underestimates variance



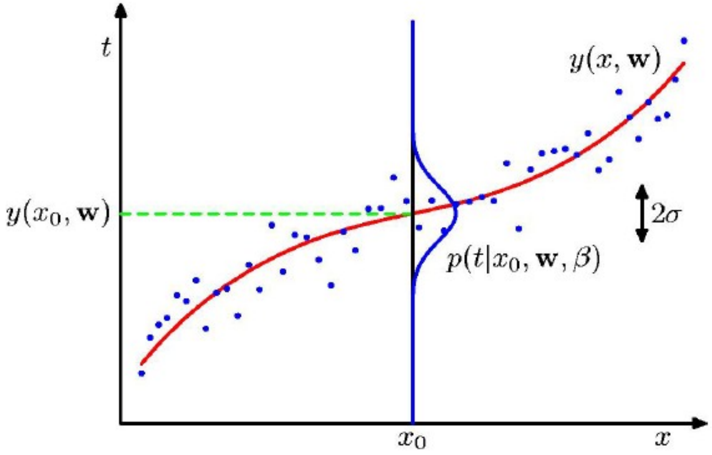
- Not an issue as increases
- Problem is related to overfitting problem

Averaged across three data sets mean is correct
Variance is underestimated because it is estimated relative

Curve Fitting (Probabilistic)

- Goal is to predict for target variable given a new value of the input variable

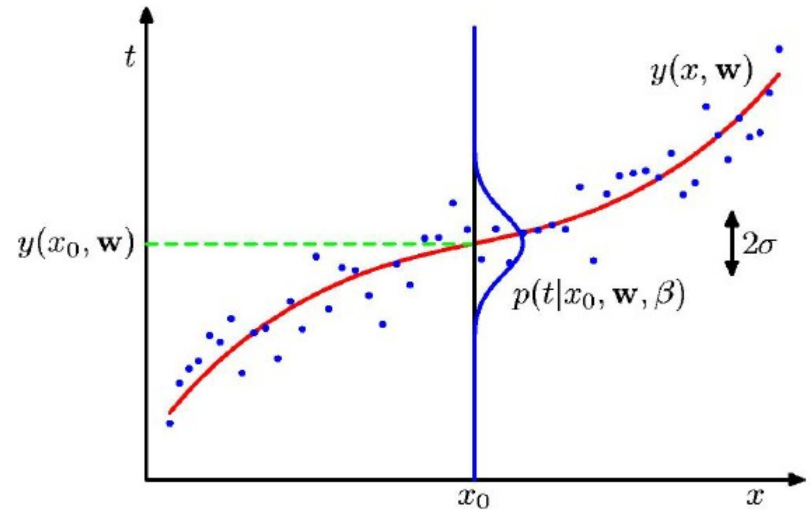
- Given input values and corresponding target values



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

Curve Fitting (Probabilistic)

- Assume given value of , value of has a Gaussian distribution with mean equal to of polynomial curve



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

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

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

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

- Logarithm of the Likelihood function is

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

- To find maximum likelihood solution for polynomial coefficients
- Maximize w.r.t.

Curve Fitting with Maximum Likelihood

- Can omit last two terms -- don't depend on
- Can replace with ½
  - since it is constant w.r.t.
- Minimize negative log-likelihood
- Identical to sum-of-squares error function

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

Precision Parameter with MLE

- Maximum likelihood can also be used to determine of Gaussian conditional distribution
- Maximizing likelihood w.r.t. gives

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

- First determine parameter vector governing the mean and subsequently use this to find precision

Predictive Distribution

- Knowing parameters and
- Predictions for new values of can be made using

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

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