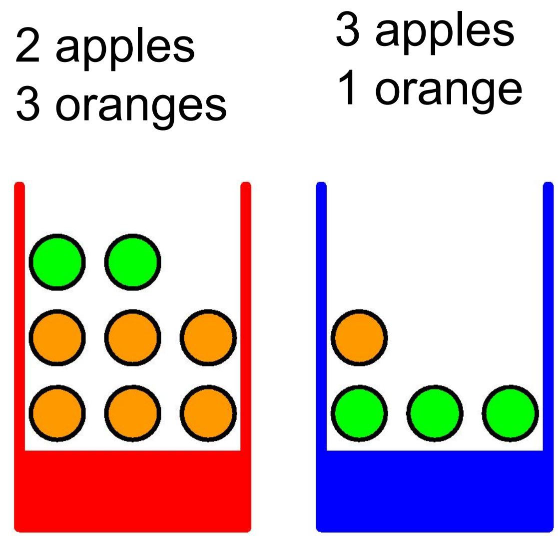
Probability Theory

Sargur N. Srihari
srihari@cedar.buffalo.edu

Probability Theory with Several Variables

- Key concept is dealing with uncertainty
  - Due to noise and finite data sets
- Framework for quantification and manipulation of uncertainty

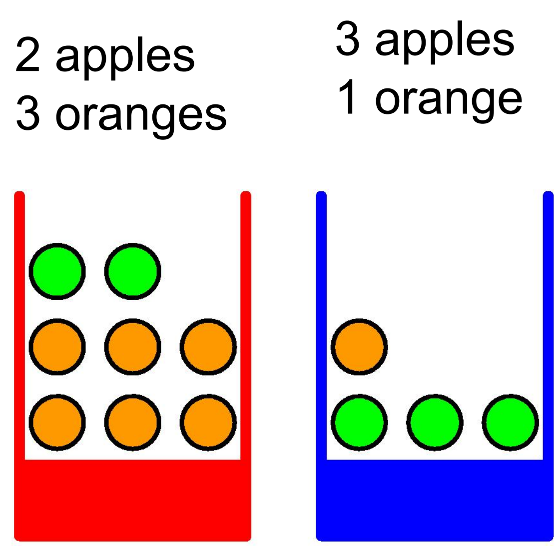


Box is random variable B (has values r or b)
Fruit is random variable F (has values o or a)

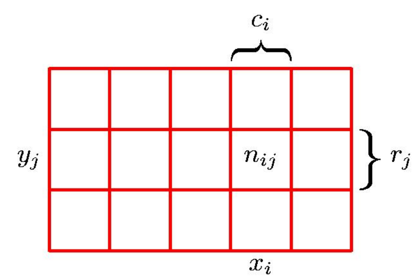
Let $p(B=r)=4/10$ and $p(B=b)=6/10$

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?



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 $Y=y_i$ is n_{ii}

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

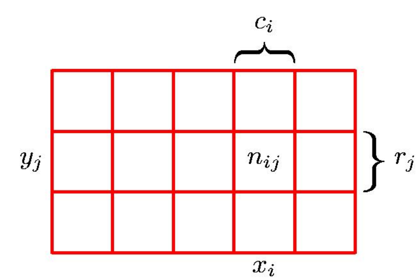
- Marginal Probability

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

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

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:



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

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

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

Bayes Theorem

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

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

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

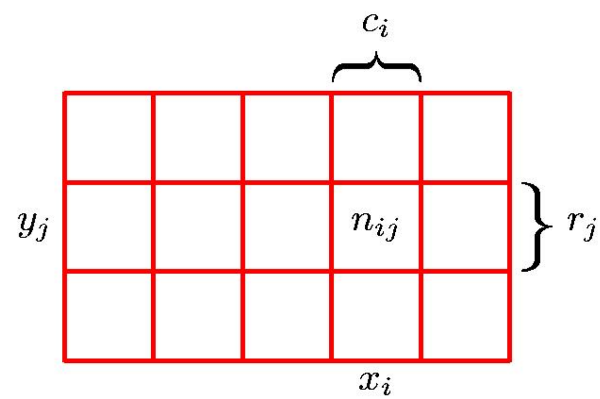
$$p(X) = \sum_{Y} p(X \mid Y)\,p(Y)$$

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

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



- 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

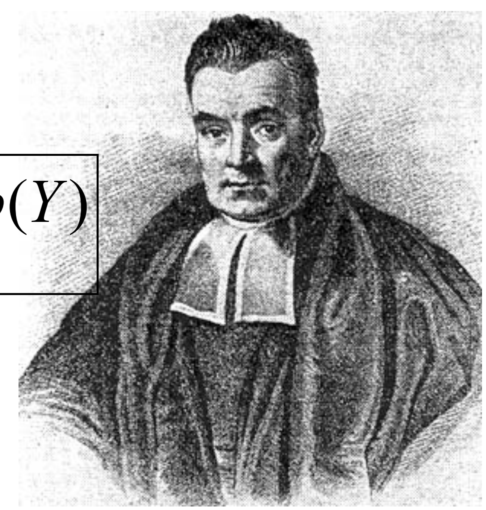
$$p(Y \mid X) = \frac{p(X \mid Y)p(Y)}{p(X)}$$

where

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

Viewed as

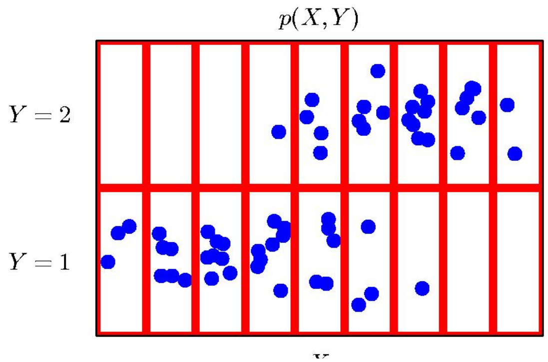
Posterior α likelihood x prior



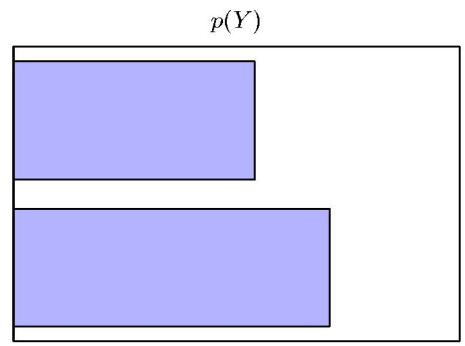
Joint Distribution over two Variables

X takes nine possible values, Y takes two values

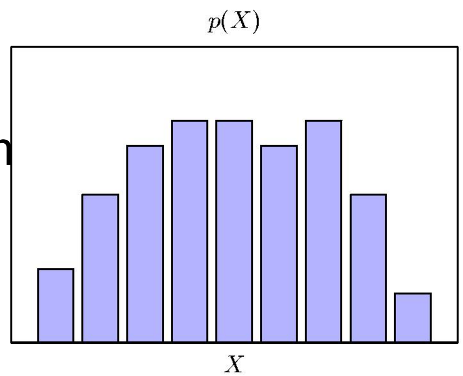
N = 60 data points



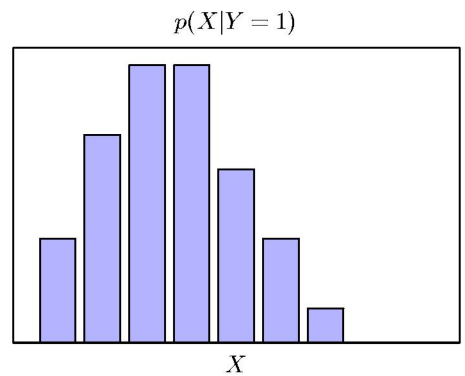
Histogram of Y (Fraction of data points having each value of Y)



Histogram of X



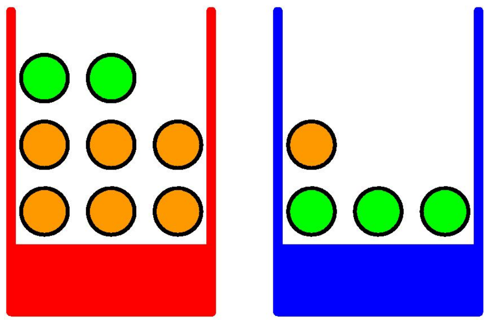
Histogram of X given Y=1



Fractions would equal the probability as N →oo

Bayes rule applied to Fruit Problem

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



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

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

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

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

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

$$= \frac{6}{8}\times\frac{4}{10} + \frac{1}{4}\times\frac{6}{10} = \boxed{\frac{9}{20} = 0.45}$$

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

Independent Variables

- If $p(X,Y)=p(X)p(Y)$ then X and Y are said to be independent
- Why?
- From product rule $$p(Y \mid X) = \frac{p(X,Y)}{p(X)} = p(Y)$$
- In fruit example if each box contained same fraction of apples and oranges then $p(F|B)=p(F)$

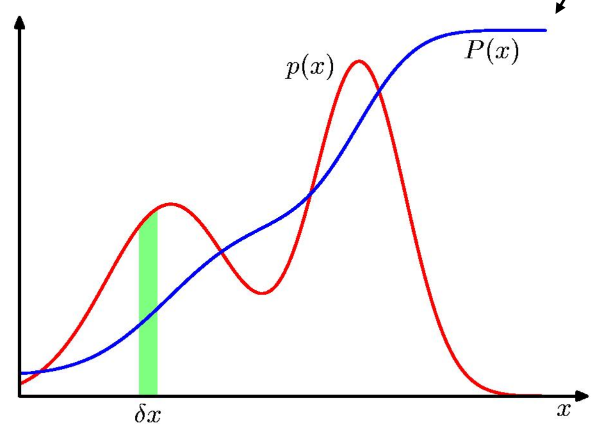
Probability Densities

- Continuous Variables
- If probability that x falls in interval $(x,x+\delta x)$ is given by $p(x)dx$ for $\delta x \to 0$

  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 x then we can define a joint probability density $p(\mathrm{x})=p(x_1,..,x_D)$
- Multivariate probability density must satisfy

$$p(\mathrm{x}) \underline{>} = 0$$

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

Sum, Product, Bayes for Continuous

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

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

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

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

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

Expectation

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

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

- 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

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

- Conditional Expectation with respect to a conditional distribution

$$E_x[f] = \Sigma_x p(x|y)\, f(x)$$

Variance

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

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

- *Expanding the square*

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

- Variance of the variable x itself

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

Covariance

- For two random variables x and y covariance is defined as

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

- 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

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

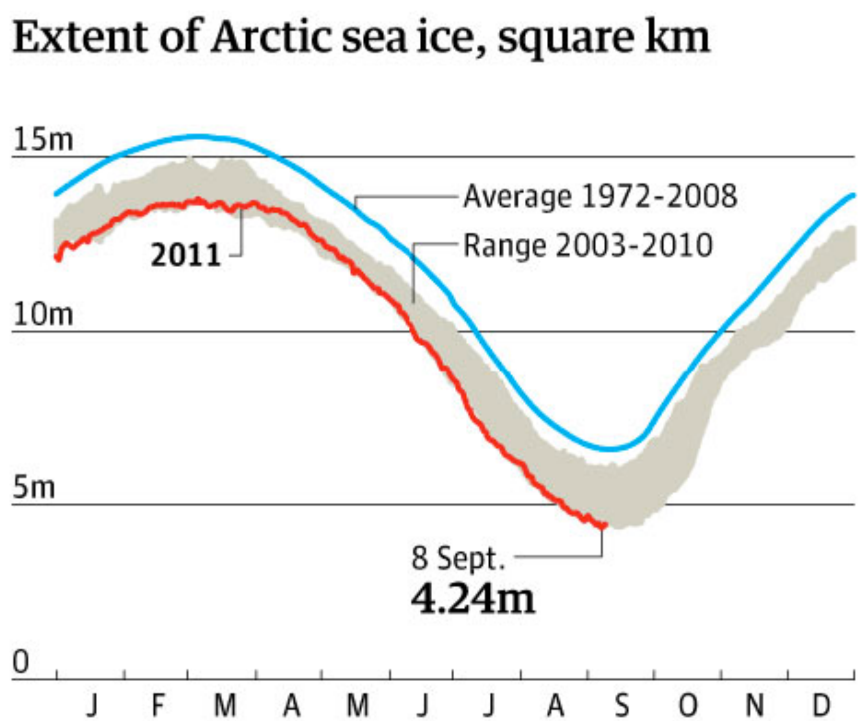
Bayesian Probabilities

- 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 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 Representation of Uncertainty

- Use of probability to represent uncertainty is not an ad-hoc choice
- If numerical values are used to represent degrees of belief, then simple set of axioms for manipulating degrees of belief leads to sum and product rules of probability (Cox's theorem)
- Probability theory can be regarded as an extension of Boolean logic to situations involving uncertainty (Jaynes)

Bayesian Approach

- Quantify uncertainty around choice of parameters $\mathbf{w}$
  - E.g., w is vector of parameters in curve fitting
- Uncertainty before observing data expressed by $p(\mathbf{w})$
- Given observed data $D = \{ t_1, \ldots t_N \}$
  - Uncertainty in w after observing D, by Bayes rule:

$$\boxed{p(\mathrm{w} \mid D) = \frac{p(D \mid \mathrm{w})\, p(\mathrm{w})}{p(D)}}$$

- Quantity $p(D|\mathbf{w})$ can be viewed as function of w
  - Represents how probable the data set is for different parameters w
    - Called Likelihood function
    - Not a probability distribution over w

Role of Likelihood Function

- Uncertainty in w expressed as

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

where $\boxed{p(D) = \int p(D \mid \mathrm{w})\, p(\mathrm{w})\, d\mathrm{w}}$ by Sum Rule

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

$$\boxed{\textit{posterior} \quad \alpha \quad \textit{likelihood} \times \textit{prior}}$$

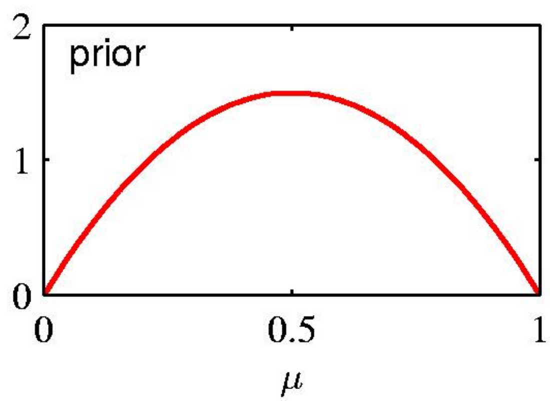
- Likelihood Function plays central role in both Bayesian and frequentist paradigms
  - Frequentist: w is a fixed parameter determined by an estimator; error bars on estimate from possible data sets D
  - Bayesian: there is a single data set D, uncertainty expressed as probability distribution over w

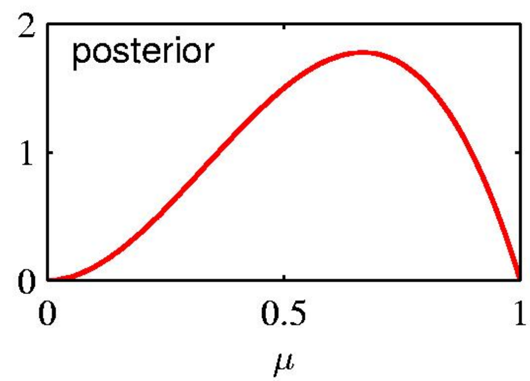
Maximum Likelihood Approach

- In frequentist setting w is considered to be a fixed parameter
  - w is set to value that maximizes likelihood function $p(D|\mathrm{w})$
  - In ML, negative log of likelihood function is called error function since maximizing likelihood is equivalent to minimizing error
  - Bootstrap approach to creating L data sets
    - From N data points new data sets are created by drawing N points at random with replacement
    - Repeat L times to generate L data sets
    - Accuracy of parameter estimate can be evaluated by variability of predictions between different bootstrap sets

Bayesian versus Frequentist Approach

- Inclusion of prior knowledge arises naturally
- Coin Toss Example
  - Fair looking coin is tossed three times and lands Head each time
  - Classical m.l.e 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





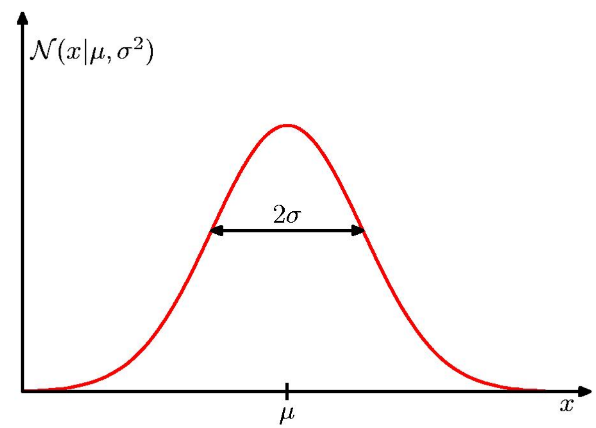
Practicality of Bayesian Approach

- 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 x

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



- Parameters:
  - Mean μ, variance σ^2,
- *Standard deviation* σ
- *Precision* $\beta = 1/\sigma^2$
- $E[x] = \mu$
- $Var[x] = \sigma^2$

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

Likelihood Function for Gaussian

- Given N observations x_i, $i=1,\ldots,n$
- Independent and identically distributed
- Probability of data set is given by likelihood function

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

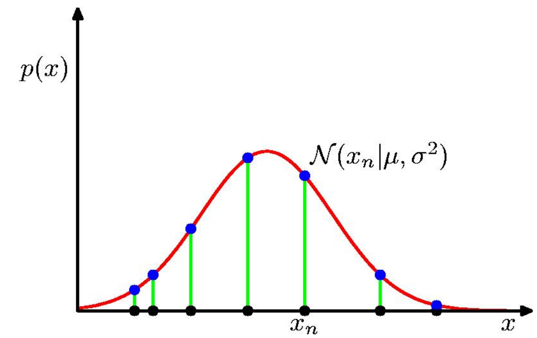
- Log-likelihood function is

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

- Maximum likelihood solutions are given by

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

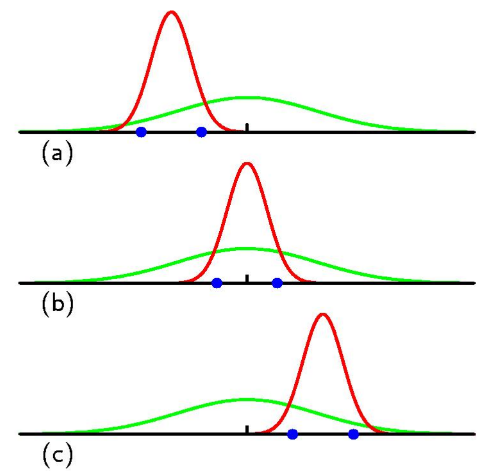
$$\sigma^2_{ML}=\frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{ML})^2$$



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
- $E[\mu_{ML}]=\mu$
- $E[\sigma^2_{ML}]=((N-1)/N)\sigma^2$



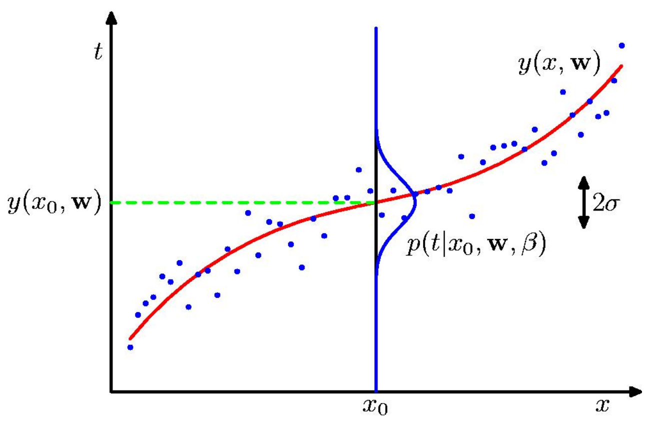
- *Problem is related to overfitting problem*

Averaged across three data sets mean is correct
Variance is underestimated because it is estimated relative to sample mean and not true mean

Curve Fitting Probabilistically

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

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



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

Curve Fitting with Maximum Likelihood

- Likelihood Function is $p(\mathrm{t} \mid \mathrm{x}, \mathrm{w}, \beta) = \prod_{n=1}^{N} N(t_n \mid y(x_n, \mathrm{w}), \beta^{-1})$
- Logarithm of the Likelihood function is

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

- To find maximum likelihood solution for polynomial coefficients w_{ML}
  - Maximize w.r.t w
  - Can omit last two terms -- don't depend on w
  - Can replace *β/2* with ½
  - Minimize negative log-likelihood
  - Identical to sum-of-squares error function

Precision parameter with Maximum Likelihood

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

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

- First determine parameter vector w_{ML} governing the mean and subsequently use this to find precision β

Predictive Distribution

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

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

- 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 $\mathbf{w}$

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

where α is the precision of the distribution

$M+1$ is the total number of parameters for an M^{th} order polynomial

(α are Model parameters also called *hyperparameters* they control distribution of model parameters)

Posterior Distribution

- Using Bayes theorem, posterior distribution for w is proportional to product of prior distribution and likelihood function

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

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

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

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

Bayesian Curve Fitting

- Previous treatment still makes point estimate of $\mathbf{w}$
  - In fully Bayesian approach consistently apply sum and product rules and integrate over all values of $\mathbf{w}$
- Given training data $\mathbf{x}$ and $\mathbf{t}$ and new test point x , goal is to predict value of t
  - *i.e,* wish to evaluate *predictive distribution* $p(t|x,\mathbf{x},\mathbf{t})$
- Applying sum and product rules of probability

$$p(t \mid x,\mathbf{x},\mathbf{t}) = \int p(t,\mathbf{w} \mid x,\mathbf{x},t)\,d\mathbf{w}$$ by Sum Rule (marginalizing over w)

$$= \int p(t \mid x,\mathbf{w},\mathbf{x},\mathbf{t})\; p(\mathbf{w} \mid x,\mathbf{x},\mathbf{t})$$ by Product Rule

$$= \int \underbrace{p(t \mid x,\mathbf{w})}_{p(t \mid x,\mathbf{w}) = N(t \mid y(x,\mathbf{w}),\beta^{-1})}\, \underbrace{p(\mathbf{w} \mid \mathbf{x},\mathbf{t})}_{\text{Posterior distribution over parameters. Also a Gaussian}}\,d\mathbf{w}$$ by eliminating unnecessary variables

Bayesian Curve Fitting

- Posterior can be shown to be Gaussian

$$p(t \mid x, \mathbf{x}, \mathbf{t}) = N(t \mid m(x), s^2(x))$$

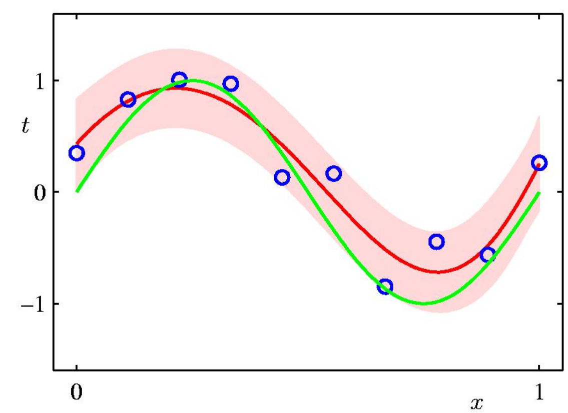
- Mean and Variance are dependent on x

$$m(x) = \beta \phi(x)^T S \sum_{n=1}^{N} \phi(x_n) t_n$$

$$s^2(x) = \beta^{-1} + \phi(x)^T S \phi(x)$$

$$S^{-1} = \alpha I + \beta \sum_{n=1}^{N} \phi(x_n) \phi(x)^T$$

$\phi(x)$ has elements $\phi_i(x) = x^i$ for $i = 0,..M$



Predictive Distribution
M=9 polynomial
α = 5 x 10^{-3}
β =11.1
Red curve is mean
Red region is $\pm$1 std dev

Model Selection

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

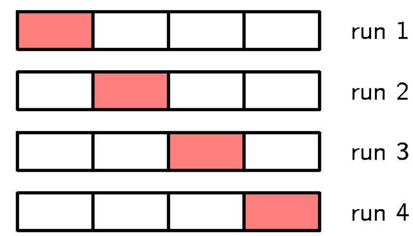
Validation Set to Select Model

- Performance on training set is not a good indicator of predictive performance
- If there is plenty of data,
  - use some of the data to train a range of models Or a given model with a range of values for its parameters
  - Compare them on an independent set, called validation set
  - Select one having best predictive performance
- If data set is small then some over-fitting can occur and it is necessary to keep aside a test set

S-fold Cross Validation

- Supply of data is limited
- All available data is partitioned into S groups
- S-1 groups are used to train and evaluated on remaining group
- Repeat for all S choices of held-out group
- Performance scores from S runs are averaged

S=4



If S=N this is the leave-one-out method

Bayesian Information Criterion

- Criterion for choosing model
- *Akaike Information criterion* (AIC) chooses model for which the quantity

$$\ln p(D|w_{ML}) - M$$

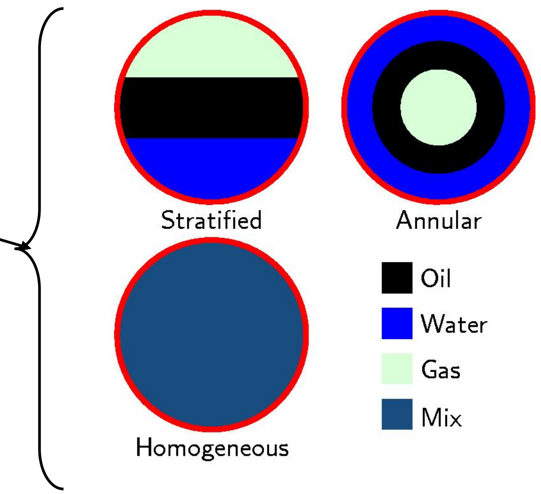
- Is highest
- Where M is number of adjustable parameters
- BIC is a variant of this quantity

The Curse of Dimensionality

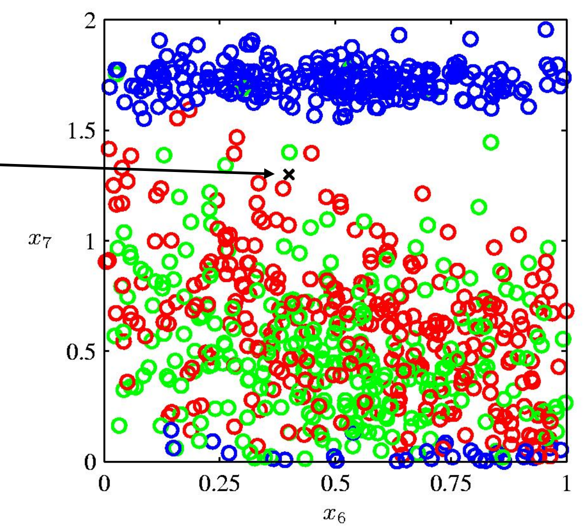
Need to deal with spaces with many variables in machine learning

Example Clasification Problem

- Three classes
- 12 variables: two shown
- 100 points
- Learn to classify from data

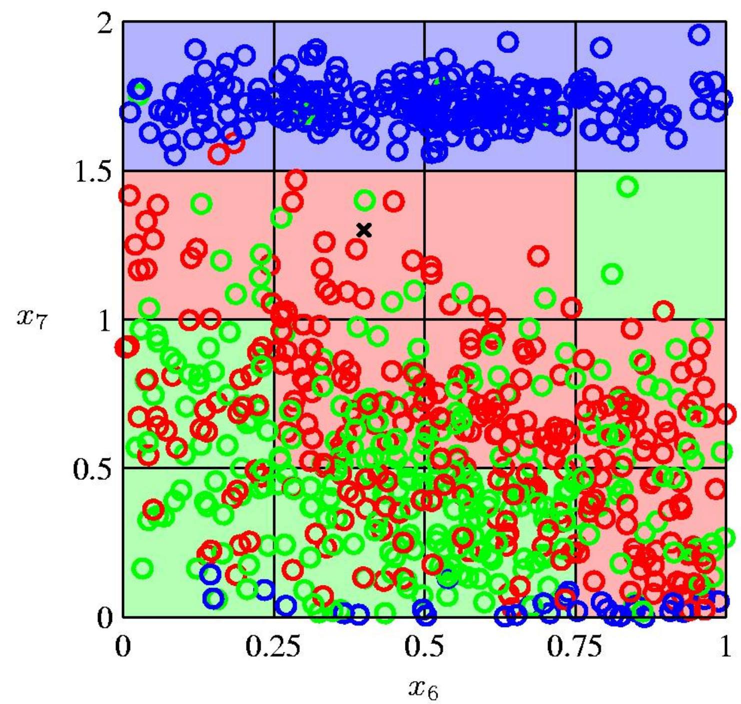


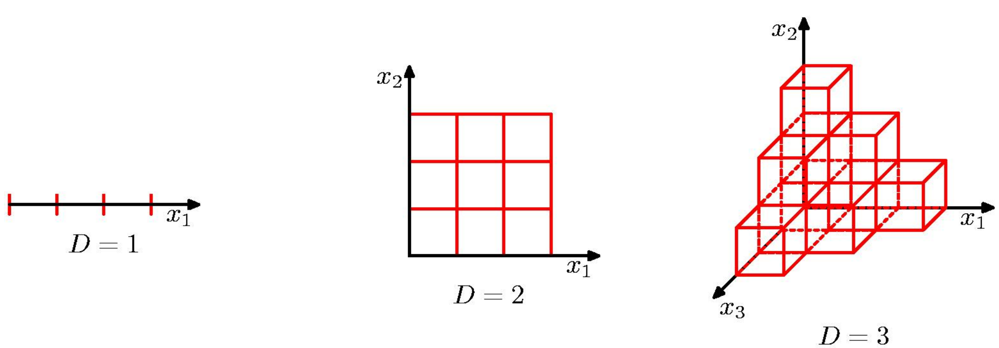
Which class should x belong to?



Cell-based Classification

- Naïve approach of cell based voting will fail because of exponential growth of cells with dimensionality
- Hardly any points in each cell

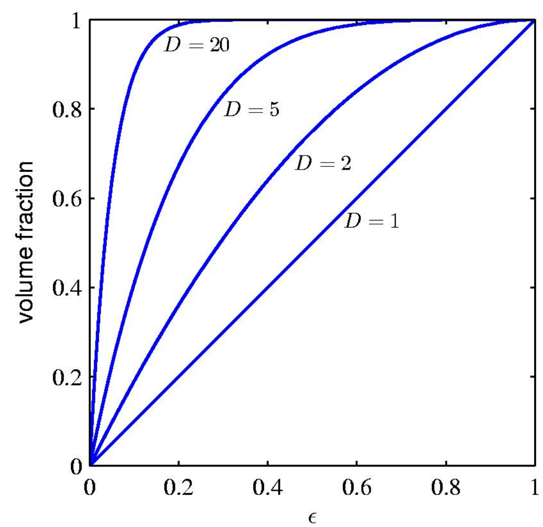




Volume of Sphere in High Dimensions

- Sphere is of radius $r=1$ in D-dimensions
- What fraction of volume lies between radius

  $r = 1-\varepsilon$ and $r=1$?
- $V_D(r)=K_D r^D$
- This fraction is given by $1-(1-\varepsilon)^D$
- As D increases high proportion of volume lies near outer shell



Fraction of volume of sphere lying in range $r = 1-\varepsilon$ to $r = 1$ for various dimensions D

Gaussian in High-dimensional Space

- x-y space converted to *r*-space using polar coordinates
- Most of the probability mass is located in a thin shell at a specific radius