Probability Theory in Machine Learning

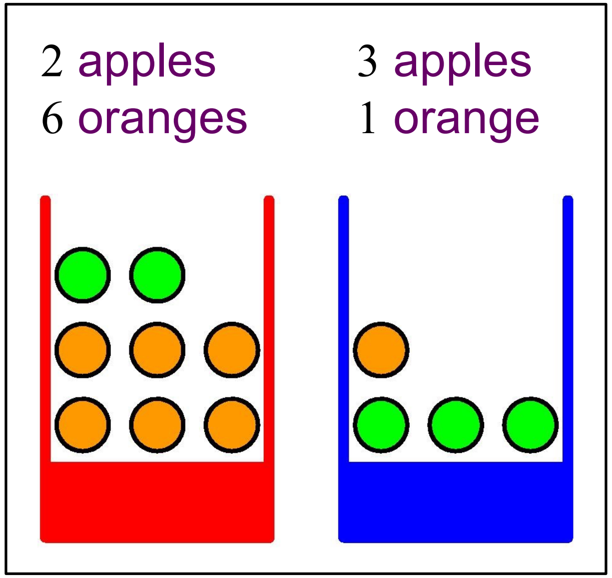
- Probability is key concept is dealing with uncertainty
  - Arises due to finite size of data sets and noise on measurements
- Probability Theory
  - Framework for quantification and manipulation of uncertainty
  - One of the central foundations of machine learning

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
  - Mathematical function that describes
    1. possible values of a r.v.
    2. and associated probabilities

Probability with Two Variables

- Key concepts:
  - conditional & joint probabilities of variables
- Random Variables: B and F
  - Box B, Fruit F
    - F has two values orange (o) or apple (a)
    - B has values red (r) or blue (b)



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

CPD: Data

Box\Fruit	*orange*	*apple*
red	6	2
blue	1	3

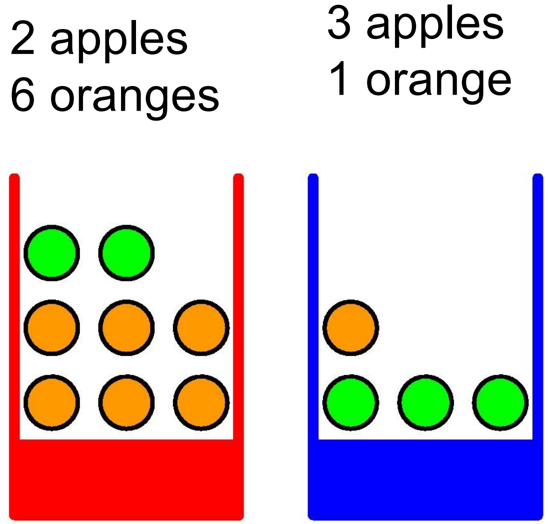
$p(F|B)$

Box\Fruit	*orange*	*apple*
red	3/4	1/4
blue	1/4	3/4

Given the above data we are interested in several probabilities of interest: *marginal, conditional and joint*

Probabilities of Interest

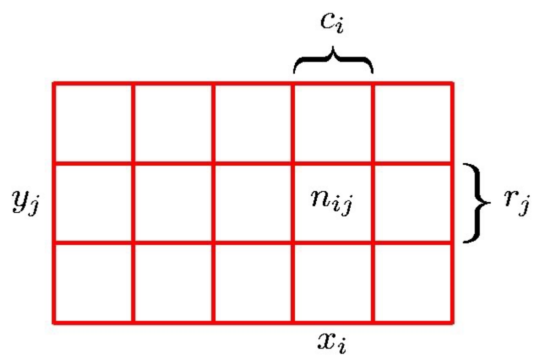
- Marginal Probability
  - what is the probability of an apple? $P(F=a)$
    - Note that we have to consider $P(B)$
- Conditional Probability
  - Given that we have an orange what is the probability that we chose the blue box? $P(B=b|F=o)$
- Joint Probability
  - What is the probability of orange AND blue box? $P(B=b,F=o)$



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

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}



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

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

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

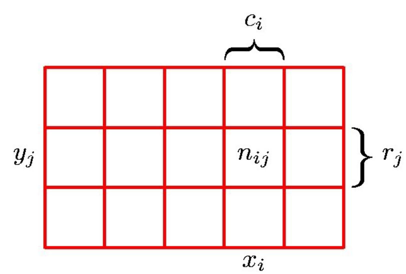
Product Rule of Probability Theory

- 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:

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

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

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



Bayes Theorem

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

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

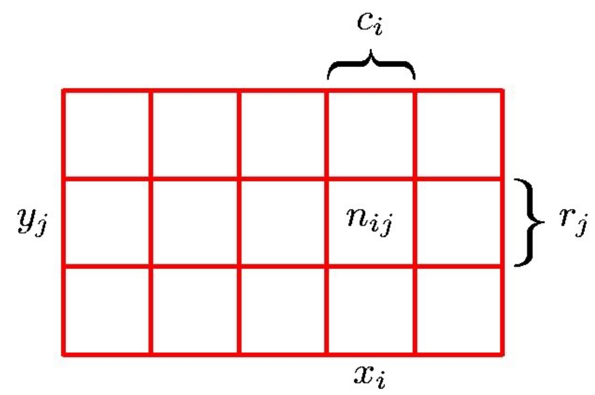
- Which is called Bayes' theorem
- Using the sum rule the denominator is expressed as

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

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

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

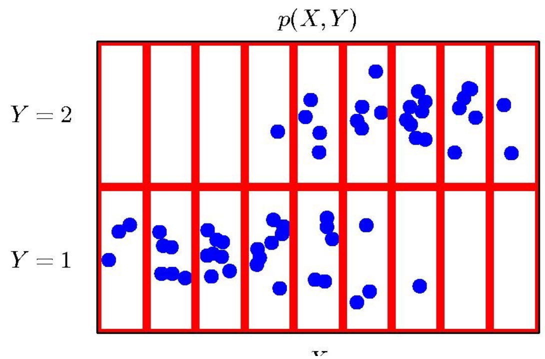
Viewed as

Posterior a likelihood x prior

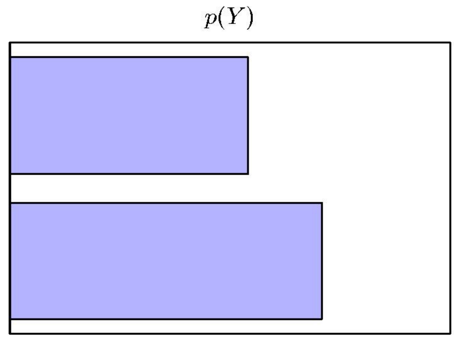
Ex: Joint Distribution over two Variables

X takes nine possible values, Y takes two values

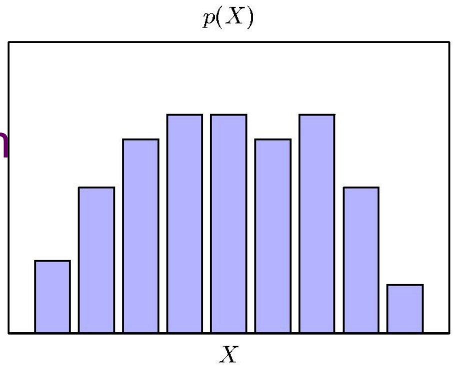
$N = 60$ data points



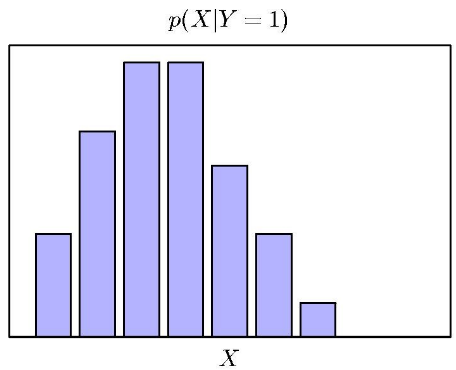
Histogram of Y (Fraction of data points having each value of Y)



Histogram of X

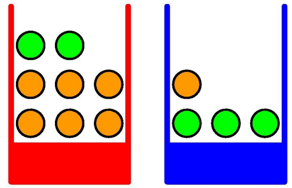


Histogram of X given Y=1



Fractions would equal the probability as $N \rightarrow \infty$

Bayes rule applied to Fruit Problem



$P(F\|B)$	*orange*	*apple*
red	3/4	1/4
blue	1/4	3/4

Priors: $p(B=r)=4/10, p(B=b)=6/10$

– Conditional Probability

- box is red given that fruit is orange

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

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

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

– Marginal Probability

- fruit is orange

$$p(F=o) = p(F=o, B=r) + p(F=o, B=b)$$ From sum rule of probability

$$= p(F=o \mid B=r)p(B=r) + p(F=o \mid B=b)p(B=b)$$ From product rule of probability

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

The *marginal* probability of 0.45 is lower than simple average 7/12=0.58, since we need box priors

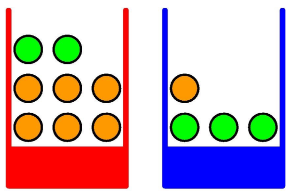
Similarly $P(F=a)=0.55$. Note that even though there are fewer apples, the blue box is more likely

$\boxed{P(B=r \mid F=a)=1/55=0.02}$

Independent and Dependent Variables

- Independent variables
  - If $p(X,Y)=p(X)p(Y)$ then X and Y are independent
  - Why?
  - From product rule: $\boxed{p(Y \mid X) = \frac{p(X,Y)}{p(X)} = p(Y)}$
- Dependent variables
  - In fruit example the variables are not independent
    - $P(F{=}o, B{=}r) = 6/8 \times 4/10 = 0.3$
    - $P(F{=}o)P(B{=}r) = 0.45 \times 0.4 = 0.18$
  - In general, $p(X,Y) > p(X)p(Y)$

Priors: $p(B{=}r){=}4/10, p(B{=}b){=}6/10$



$P(F\|B)$	*orange*	*apple*
red	3/4	1/4
blue	1/4	3/4

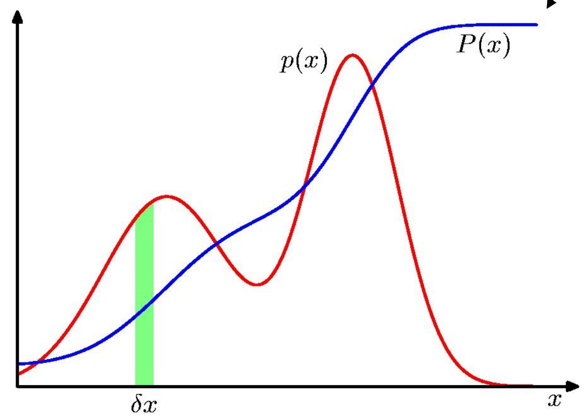
Probability Density Function (pdf)

- Continuous Variables
- If probability that x falls in interval $(x, x + \delta x)$ is given by $p(x)dx$ 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 x then we can define a joint probability density $p(\mathrm{x})=p(x_1,..,x_D)$
- Multivariate probability density must satisfy

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

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

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

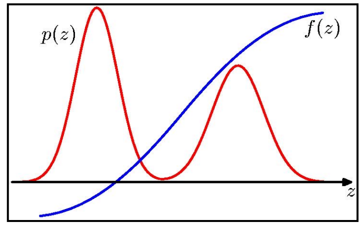
Expectation

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

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

- For a continuous distribution

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



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

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

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

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

- Conditional Expectation with respect to a conditional distribution

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

Variance

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

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

- *Expanding the square*

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

- Variance of the variable x itself

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

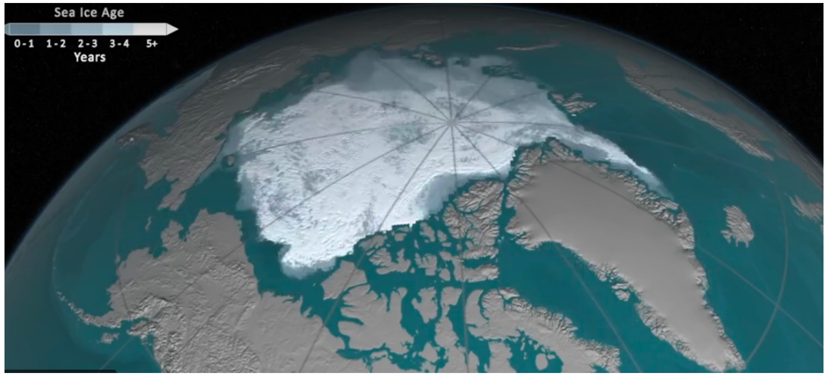
Covariance

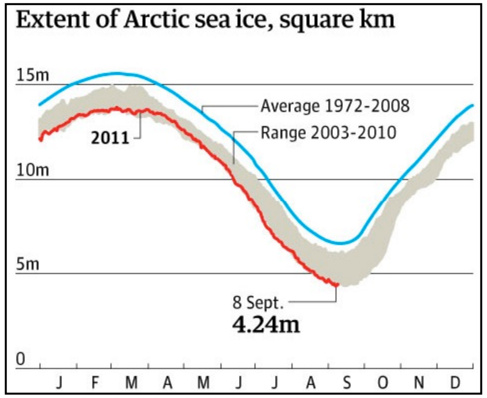
- For two random variables x and y their covariance is
- $$\mathrm{cov}[x,y] = E_{x,y}[\{x-E[x]\}\{y-E[y]\}]$$
  $$= E_{x,y}[xy] - E[x]E[y]$$
  - Expresses how x and y vary together
- If x and y are independent then their covariance vanishes
- If x and y are two vectors of random variables covariance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

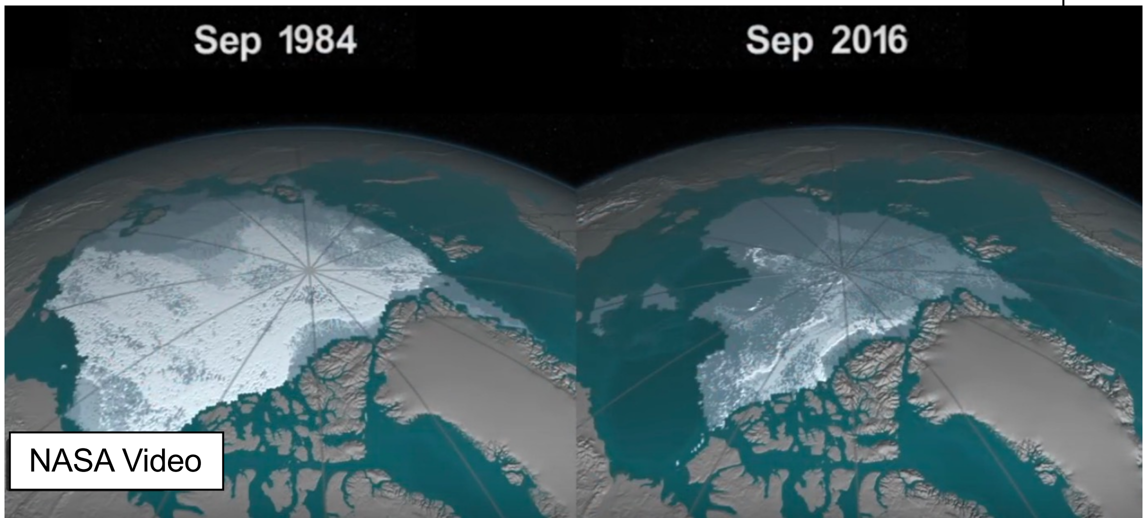
Bayesian Probabilities

- 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

Whether Arctic Sea cap will disappear







- 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)

An uncertain event
Answered by general Bayesian interpretation

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

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

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

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

  - 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 × prior

- Denominator is normalization factor
  - Involves marginalization over w

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

Role of Likelihood Function

- Likelihood Function plays central role in both *Bayesian* and *frequentist* paradigms
- Frequentist:
  - $\mathbf{w}$ is a fixed parameter determined by an estimator;
  - Error bars on estimate are obtained from possible data sets D
- Bayesian:
  - There is a single data set D
  - Uncertainty in parameters expressed as probability distribution over $\mathbf{w}$

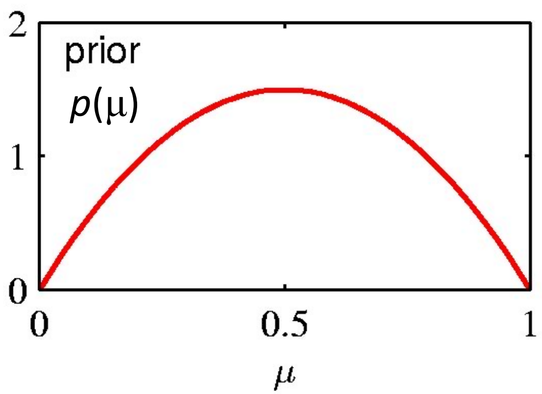
Maximum Likelihood Approach

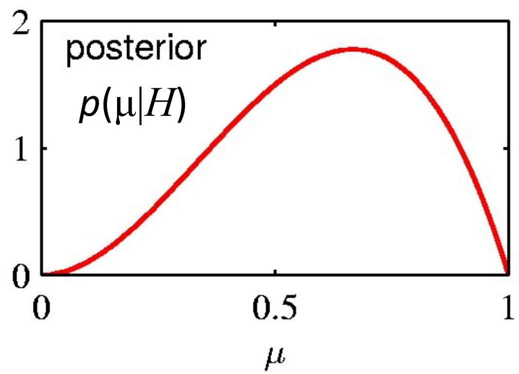
- In frequentist setting w is a fixed parameter
  - w is set to value that maximizes likelihood function $p(D|\mathrm{w})$
  - In ML, negative log of likelihood function is called error function since maximizing likelihood is equivalent to minimizing error
- Error Bars
  - Bootstrap approach to creating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

$\mu = p(H)$





Practicality of Bayesian Approach

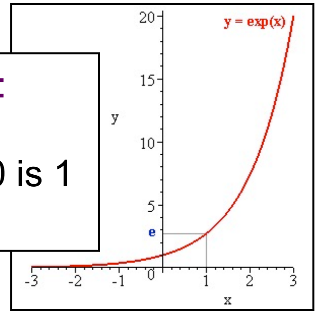
- Marginalization over whole parameter space is required to make predictions or compare models
- Factors making it practical:
  - Sampling Methods such as *Markov Chain Monte Carlo* methods
  - Increased speed and memory of computers
- Deterministic approximation schemes such as *Variational Bayes* and *Expectation propagation* are alternatives to sampling

The Gaussian Distribution

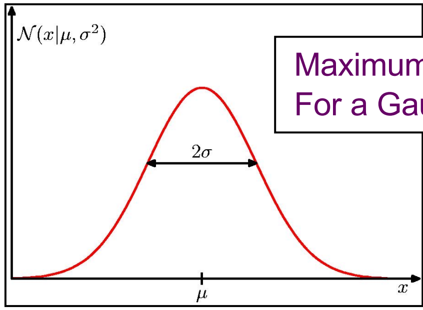
- For single real-valued variable *x*

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

What is an Exponential:
$y=e^x$, where e=2.718
Its value for argument 0 is 1
It is its own derivative



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



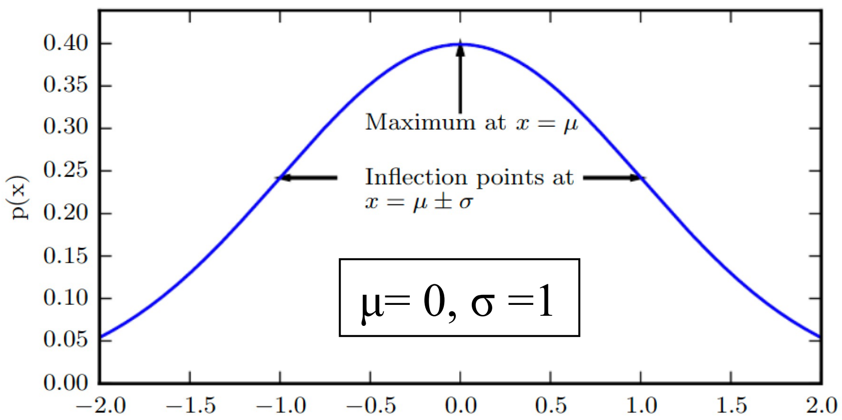
Maximum of a distribution is its mode
For a Gaussian, mode coincides with mean

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

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

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

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

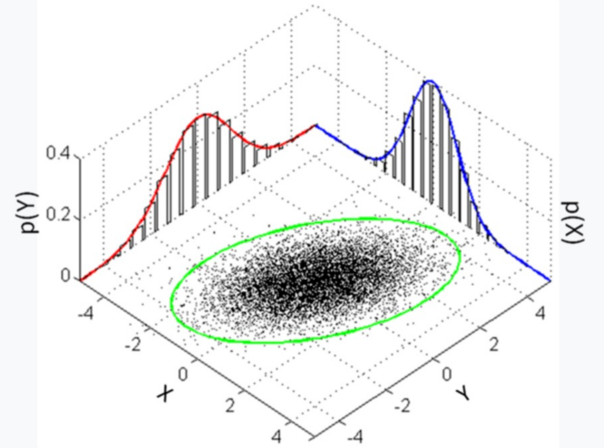


Multivariate Gaussian Distribution

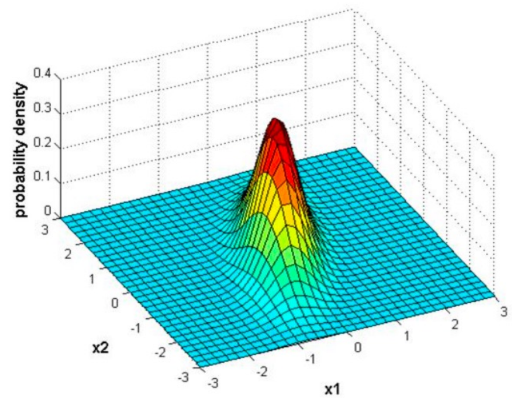
- For single real-valued variable x

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

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



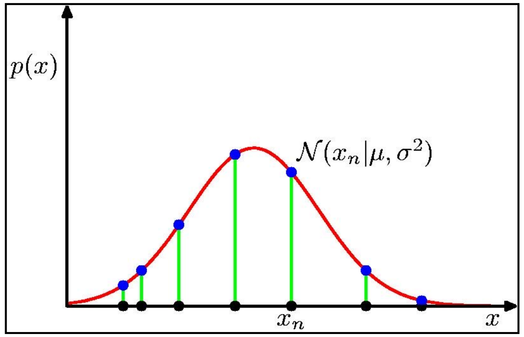
Many sample points from a multivariate normal distribution with $\boldsymbol{\mu}=\begin{bmatrix}0\\0\end{bmatrix}$ and $\boldsymbol{\Sigma}=\begin{bmatrix}1 & 3/5\\ 3/5 & 2\end{bmatrix}$, shown along with the 3-sigma ellipse, the two marginal distributions, and the two 1-d histograms.



$$\boldsymbol{\mu}=\begin{bmatrix}0\\0\end{bmatrix}$$

$$\boldsymbol{\Sigma}=\begin{bmatrix}0.25 & 0.3\\ 0.3 & 1\end{bmatrix}$$

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

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

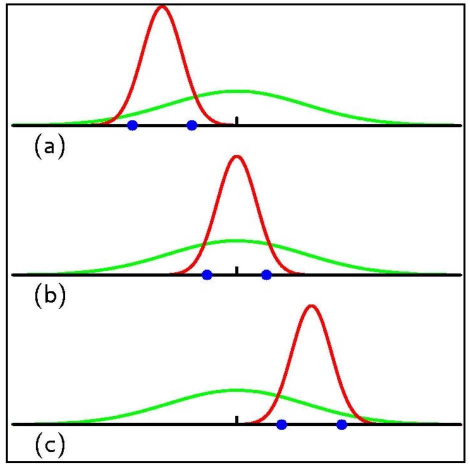
- Maximum likelihood solutions are given by

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

Bias in Maximum Likelihood

- Maximum likelihood systematically underestimates variance
  - $E[\mu_{ML}]=\mu$
  - $E[\sigma^2_{ML}]=((N-1)/N)\sigma^2$
  - Not an issue as N increases
- Problem is related to *over-fitting* problem



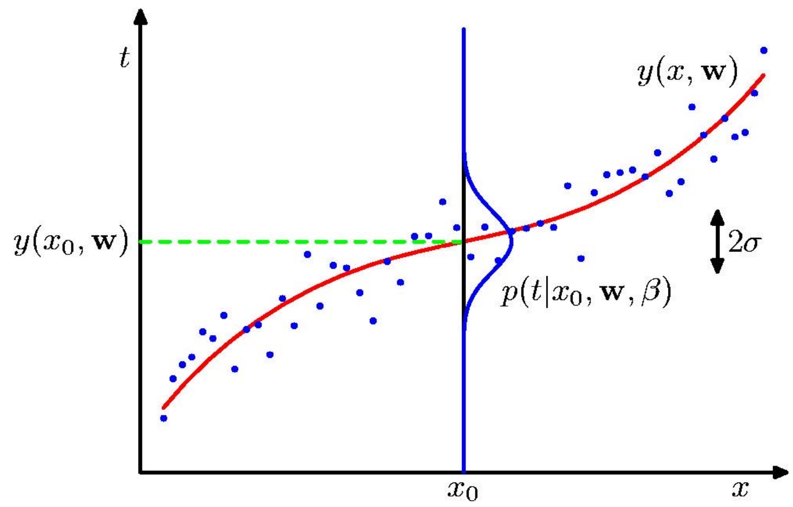
Green curve is true distribution
Averaged across three data sets mean is correct
Variance is underestimated because it is estimated relative to sample mean and not true mean

Curve Fitting Probabilistically

- Goal is to predict for target variable t given a new value of the input variable x
  - Given N input values $\mathbf{x}=(x_1,..x_N)^T$ 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 **w**
  - Can omit last two terms -- don't depend on **w**
  - Can replace β/2 with ½ (since it is constant wrt **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

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

- $\mathbf{w}$ 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 $\lambda=\alpha/\beta$

Bayesian Curve Fitting

- Previous treatment still makes point estimate of **w**
  - In fully Bayesian approach consistently apply sum and product rules and integrate over all values of **w**
- Given training data **x** and **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
  - Predictive distribution can be written in the form

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

$$= \int p(t|x, \mathbf{w}, \mathbf{x}, \mathbf{t})\ p(\mathrm{w} \mid x, \mathrm{x}, \mathbf{t}) \qquad \text{by Product Rule}$$

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

Bayesian Curve Fitting

- Predictive distribution is also Gaussian

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

  – Where the Mean and Variance are dependent on *x*

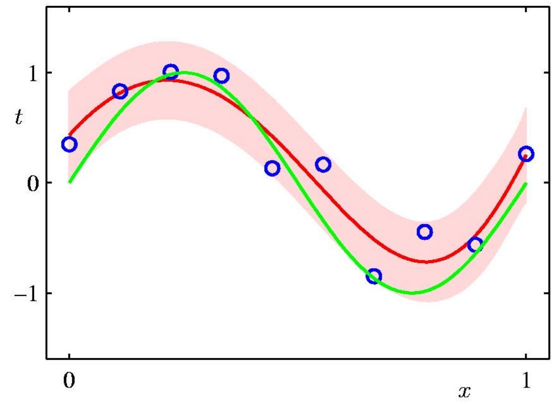
$$m(x) = \beta \phi(x)^T S \sum_{n=1}^{N} \phi(x_n) t_n$$

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

First term is uncertainty in predicted value due to noise in target
Second term is uncertainty in parameters due to Bayesian treatment

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

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



Predictive Distribution is a 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

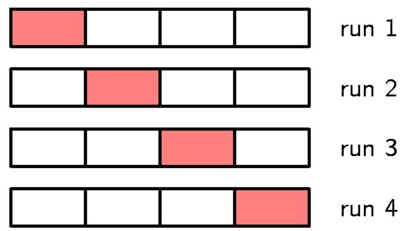
Validation Set to Select Model

- 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

$$\ln p(\mathrm{D}|\mathrm{w}_{\mathrm{ML}}) - \mathrm{M}$$

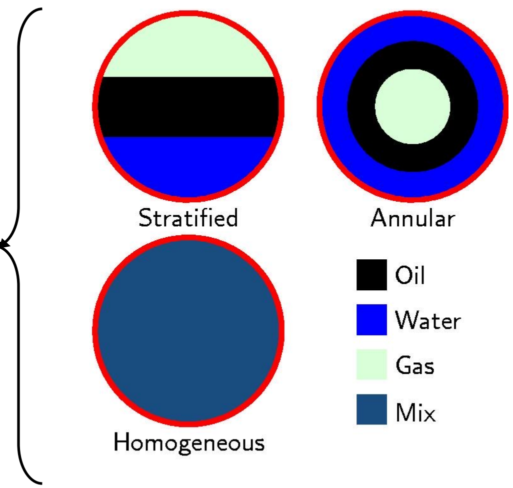
- Is highest
- Where M is number of adjustable parameters
- BIC is a variant of this quantity

The Curse of Dimensionality

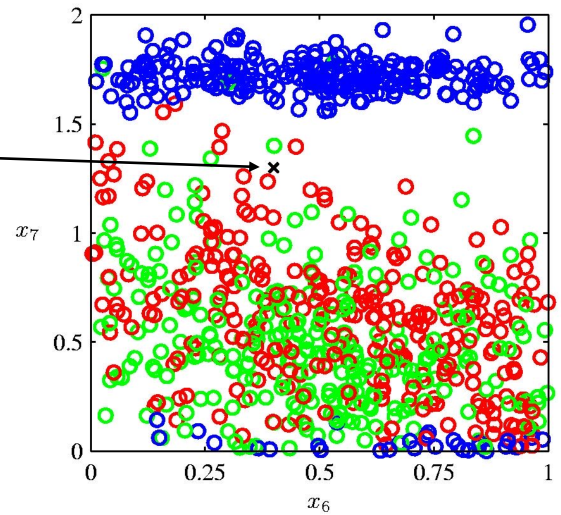
Need to deal with spaces with many variables in machine learning

Example Clasification Problem

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

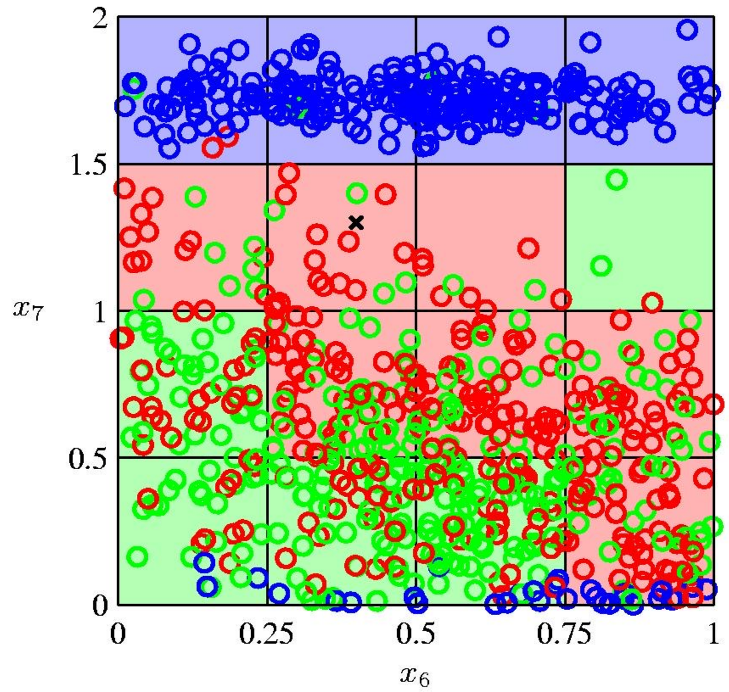


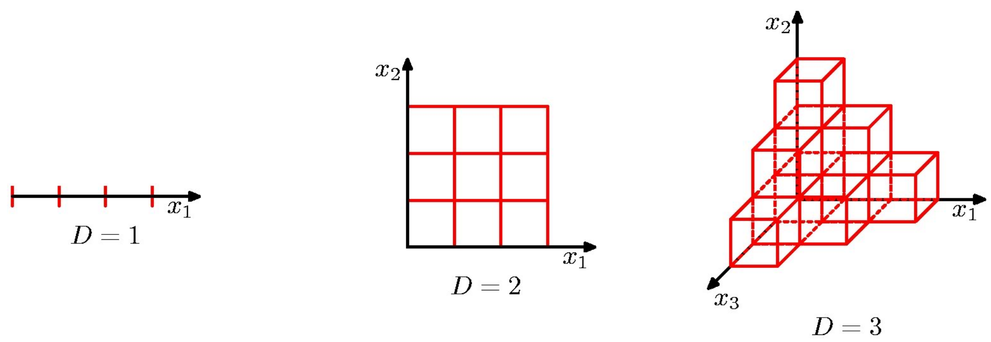
Which class should x belong to?



Cell-based Classification

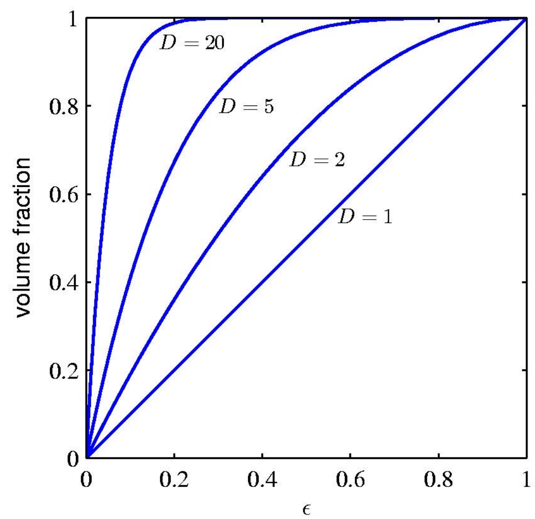
- Naïve approach of cell based voting will fail because of exponential growth of cells with dimensionality
- Hardly any points in each cell





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