Statistical Machine Learning

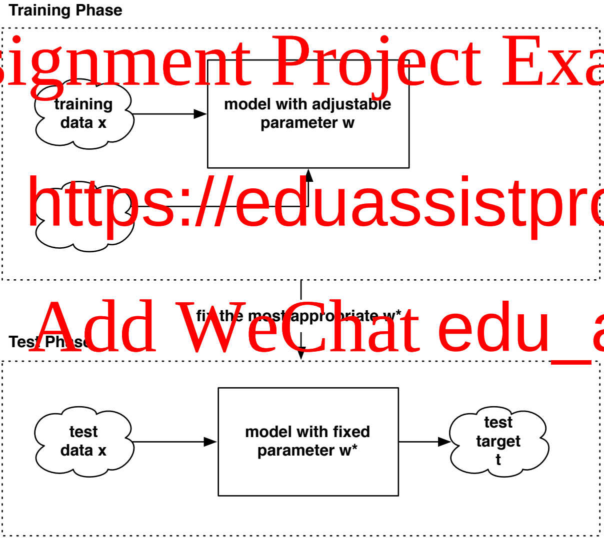
Canberra
Semester One, 2021.

(Many figures from C. M. Bishop, "Pattern Recognition and Machine Learning")

Part I

- Bas
- Ma
- Re
- Bias variance decomposition

Training and Testing: (Non-Bayesian) Point Estimate



Ong & Walder & Webers & Xie
Data61 | CSIRO
ANU Computer Science

Bayesian Regression

- Bayes Theorem

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{} \qquad p(\mathbf{w} \mid \mathbf{t}) = \frac{p(\mathbf{t} \mid \mathbf{w})\, p(\mathbf{w})}{}$$

wh

and

- I.i.d

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

$$= \prod_{n=1}^{N} \mathcal{N}(t_n \mid \mathbf{w}^\top \phi(\mathbf{x}_n), \beta^{-1})$$

$$= \text{const} \times \exp\{-\beta \frac{1}{2} (\mathbf{t} - \mathbf{\Phi} \mathbf{w})^\top (\mathbf{t} - \mathbf{\Phi} \mathbf{w})\}$$

$$= \mathcal{N}(\mathbf{t} \mid \mathbf{\Phi} \mathbf{w}, \beta^{-1} \mathbf{I})$$

- The choice of prior affords an intuitive control over our inductive bias
- All inference schemes have such biases, and often arise mor
- Ca
  -
  -

An answer to the second question:

Definition (Conjugate Prior)

A class of prior probability distributions $p(w$
class of likelihood functions $p(x \mid w)$ if the resulting posterior distributions $p(w \mid x)$ are in the same family as $p(w)$.

Examples of Conjugate Prior Distributions

Table: Discrete likelihood distributions

Likelihood	Conjugate Prior
Bernoulli	Beta

Table: Continuous likelihood distribu

Likelihood	Conjuga
Uniform	Pareto
Exponential	Gamma
Normal	Normal (mean parameter)
Multivariate normal	Multivariate normal (mean parameter)

Conjugate Prior to a Gaussian Distribution

- Example : If the likelihood function is Gaussian, choosing a Gaussian prior for the mean will ensure that the posterior distribution is also Gaussian.
- Given a marginal distribution for $\mathbf{x}$ and a conditional Gaussian distribution for given in the form

- we get

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}\,|\,\boldsymbol{A}\boldsymbol{\mu} + \mathbf{b}, \boldsymbol{L}^{-1} +$$

$$p(\mathbf{x}\,|\,\mathbf{y}) = \mathcal{N}(\mathbf{x}\,|\,\boldsymbol{\Sigma}\{\boldsymbol{A}^\top \boldsymbol{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma})$$

where $\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \boldsymbol{A}^\top \boldsymbol{L} \boldsymbol{A})^{-1}$.

Note that the covariance $\boldsymbol{\Sigma}$ does not involve $\boldsymbol{y}$.

Conjugate Prior to a Gaussian Distribution (intuition)

Given

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}\,|\,\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$$

$$p(\mathbf{y}\,|\,\mathbf{x}) = \mathcal{N}(\mathbf{y}\,|\,A\mathbf{x}+\mathbf{b}, L^{-1}) \Leftrightarrow \boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{b} + \mathcal{N}(\mathbf{0}, L^{-1})$$

We have $\mathbb{E}[\boldsymbol{y}] = \mathbb{E}[A\boldsymbol{x}+\boldsymbol{b}] = A\boldsymbol{\mu}+\boldsymbol{b}$ and by the easily proven Bienaymé formula for t

$$\mathrm{co} \qquad \mathrm{cov}[$$

So $\boldsymbol{y}$ is Gau

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}\,|\,A\boldsymbol{\mu}+\mathbf{b}, L^{-1} + A\boldsymbol{\Lambda}^{-}$$

Then letting $\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + A^\top L A)^{-1}$ and

$$p(\mathbf{x}\,|\,\mathbf{y}) = \mathcal{N}(\mathbf{x}\,|\,\boldsymbol{\Sigma}\{A^\top L(\mathbf{y}-\mathbf{b}) + \boldsymbol{\Lambda}$$

$$\Leftrightarrow \boldsymbol{x} = \boldsymbol{\Sigma}\{A^\top L(\mathbf{y}-\mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\} + \mathcal{N}(\mathbf{0}, \Sigma)$$

yields the correct moments for $\boldsymbol{x}$, since

$$\mathbb{E}[\boldsymbol{x}] = \mathbb{E}[\boldsymbol{\Sigma}\{A^\top L(\mathbf{y}-\mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\}] = \boldsymbol{\Sigma}\{A^\top L(A\boldsymbol{\mu}+\mathbf{b}-\mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\}$$

$$= \boldsymbol{\Sigma}\{A^\top LA\boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{\mu}\} = (\boldsymbol{\Lambda} + A^\top LA)^{-1}\{A^\top LA + \boldsymbol{\Lambda}\}\boldsymbol{\mu} = \boldsymbol{\mu},$$

and it is similar (but tedious ; don't do it) to recover $\mathrm{cov}[\boldsymbol{x}] = \Lambda$.

- Choose a Gaussian prior with mean $\mathbf{m}_0$ and covariance $\mathbf{S}_0$

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}\,|\,\mathbf{m}_0, \mathbf{S}_0)$$

- Same likelihood as before (here written in vector form):

- Giv

$$p(\mathbf{w}\,|\,\mathbf{t}) = \mathcal{N}(\mathbf{w}\,|\,\mathbf{m}_N, \mathbf{S}$$

where

$$\mathbf{m}_N = \mathbf{S}_N(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\boldsymbol{\Phi}^\top\mathbf{t})$$
$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}^\top\boldsymbol{\Phi}$$

(derive this with the identities on the previous slides)

Bayesian Regression: Zero Mean, Isotropic Prior

- For simplicity we proceed with $\mathbf{m}_0 = 0$ and $\mathbf{S}_0 = \alpha^{-1}\mathbf{I}$, so

$$p(\mathbf{w} \mid \alpha) = \mathcal{N}(\mathbf{w} \mid 0, \alpha^{-1}\mathbf{I})$$

- The posterior becomes $p(\mathbf{w} \mid \mathbf{t}) = \mathcal{N}(\mathbf{w} \mid \mathbf{m}_N, \mathbf{S}_N)$ with

- For $\alpha \ll \beta$ we get

$$\mathbf{m}_N \to \mathbf{w}_{ML} = (\mathbf{\Phi}^\top \mathbf{\Phi})^{-1}$$

- Log of posterior is sum of log likelihood and log of prior

$$\ln p(\mathbf{w} \mid \mathbf{t}) = -\frac{\beta}{2}(\mathbf{t} - \mathbf{\Phi}\mathbf{w})^\top(\mathbf{t} - \mathbf{\Phi}\mathbf{w}) - \frac{\alpha}{2}\mathbf{w}^\top\mathbf{w} + \text{const}$$

Bayesian Regression

- Log of posterior is sum of log likelihood and log of prior

$$\ln p(\mathbf{w} \mid \mathbf{t}) = -\beta \underbrace{\frac{1}{2}\|\mathbf{t} - \Phi\mathbf{w}\|^2}_{\text{sum-of-squares-error}} - \underbrace{\frac{\alpha}{2}\|\mathbf{w}\|^2}_{\text{regulariser}} + \text{const.}$$

- The

corresponds to minimising the sum-of-squa
function with quadratic regularisation coeffic

- The posterior is Gaussian so mode $=$
- For $\alpha \ll \beta$ the we recover unregularised least squares (equivalently m.a.p. approaches maximum likelihood), for example in case of
  - an infinitely broad prior with $\alpha \to 0$
  - an infinitely precise likelihood with $\beta \to \infty$

Bayesian Inference in General: Sequential Update of Belief

- If we have not yet seen any data point ($N = 0$), the posterior is equal to the prior.

- Sequential arrival of data points : the posterior given some observed data acts as the prior for the future data.
- Nicely fits a sequential learning framework.

Sequential Update of the Posterior

- Example of a linear basis function model
- Single input x, single output t
- Lin
- Tru

1.
2. Calculate $f(x_n, \mathbf{a}) = a_0 + a_1 x_n$, where $a_0 = \quad 0.3$, $a_1 = 0.5$.
3. Add Gaussian noise with standard deviation

$$t_n = \mathcal{N}(x_n \mid f(x_n, \mathbf{a}), 0$$

- Set the precision of the uniform prior to $\alpha = 2.0$.

Predictive Distribution

- In the training phase, data $\mathbf{x}$ and targets $\mathbf{t}$ are provided
- In the test phase, a new data value x is given and the corr
- Ba
  giv
  targets $\mathbf{t}$

$$p(t \mid x, \mathbf{x}, \mathbf{t})$$

- This is the Predictive Distribution (c.f. t distribution, which is over the parameters).

How to calculate the Predictive Distribution?

- Introduce the model parameter $\mathbf{w}$ via the sum rule

$$p(t \mid x, \mathbf{x}, \mathbf{t}) = \int p(t, \mathbf{w} \mid x, \mathbf{x}, \mathbf{t}) d\mathbf{w}$$

$$= \quad p(t \mid \mathbf{w}, x, \mathbf{x}, \mathbf{t}) p(\mathbf{w} \mid x, \mathbf{x}, \mathbf{t}) d\mathbf{w}$$

- The t
mo
trai

$$p(t \mid \mathbf{w}, x, \mathbf{x}, \mathbf{t}) = p(t \mid \mathbf{w}$$

- The model parameter $\mathbf{w}$ are learned with t
$\mathbf{x}$ and the training targets $\mathbf{t}$ only

$$p(\mathbf{w} \mid x, \mathbf{x}, \mathbf{t}) = p(\mathbf{w} \mid \mathbf{x}, \mathbf{t})$$

- Predictive Distribution

$$p(t \mid x, \mathbf{x}, \mathbf{t}) = \int p(t \mid \mathbf{w}, x) p(\mathbf{w} \mid \mathbf{x}, \mathbf{t}) d\mathbf{w}$$

Proof of the Predictive Distribution

The predictive distribution is

$$p(t \mid x, \mathbf{x}, \mathbf{t}) = \int p(t \mid \mathbf{w}, x, \mathbf{x}, \mathbf{t}) p(\mathbf{w} \mid x, \mathbf{x}, \mathbf{t}) d\mathbf{w}$$

because

$$\int p(t \mid$$

$$= \frac{p(t, x, \mathbf{x}, \mathbf{t})}{p(x, \mathbf{x}, \mathbf{t})}$$

$$= p(t \mid x, \mathbf{x}, \mathbf{t}),$$

or simply

$$\int p(t \mid \mathbf{w}, x, \mathbf{x}, \mathbf{t}) p(\mathbf{w} \mid x, \mathbf{x}, \mathbf{t}) d\mathbf{w} = \int p(t, \mathbf{w} \mid x, \mathbf{x}, \mathbf{t}) d\mathbf{w}$$
$$= p(t \mid x, \mathbf{x}, \mathbf{t}).$$

Predictive Distribution with Simplified Prior

- Find the predictive distribution

$$p(t\,|\,\mathbf{t},\alpha,\beta) = \int p(t\,|\,\mathbf{w},\beta)\,p(\mathbf{w}\,|\,\mathbf{t},\alpha,\beta)\,d\mathbf{w}$$

(remember : conditioning on $\mathbf{x}$ is often suppressed to sim

- No
on

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

- and the posterior was

$$p(\mathbf{w}\,|\,\mathbf{t},\alpha,\beta) = \mathcal{N}(\mathbf{w}\,|\,\mathbf{m}_N, \mathbf{S}_N)$$

where

$$\mathbf{m}_N = \beta\mathbf{S}_N\mathbf{\Phi}^\top\mathbf{t}$$
$$\mathbf{S}_N^{-1} = \alpha\mathbf{I} + \beta\mathbf{\Phi}^\top\mathbf{\Phi}$$

Predictive Distribution with Simplified Prior

©2021
Ong & Walder & Webers & Xie
Data61 | CSIRO
ANU Computer Science

- If we do the integral (it turns out to be the convolution of the two Gaussians), we get for the predictive distribution

$$p(t \mid \mathbf{x}, \mathbf{t}, \alpha, \beta) = \mathcal{N}(t \mid \mathbf{m}_N^\top \phi(\mathbf{x}), \sigma_N^2(\mathbf{x}))$$

wh

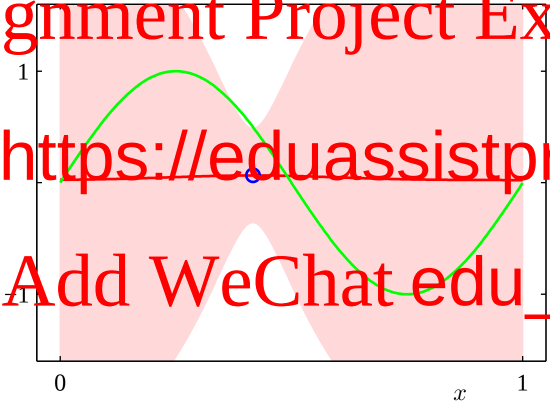
- This is more easily shown using a similar approa earlier "intuition" slide and again with the Bienay formula, now using

$$t = \boldsymbol{w}^\top \phi(\boldsymbol{x}) + \mathcal{N}(0, \beta^{-1}).$$

However this is a linear-Gaussian specific trick and in general we need to integrate out the parameters.

Predictive Distribution with Simplified Prior

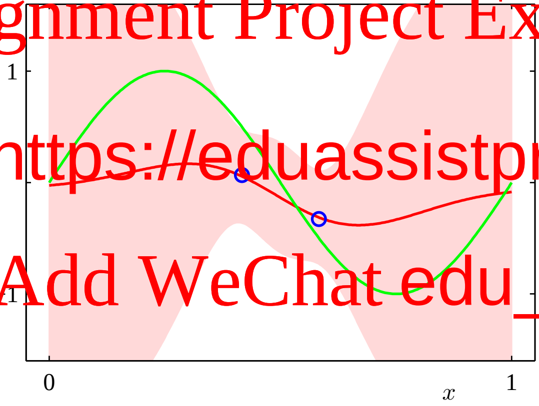
Example with artificial sinusoidal data from $\sin(2\pi x)$ (green) and added noise. Number of data points $N = 1$.



Mean of the predictive distribution (red) and regions of one standard deviation from mean (red shaded).

Predictive Distribution with Simplified Prior

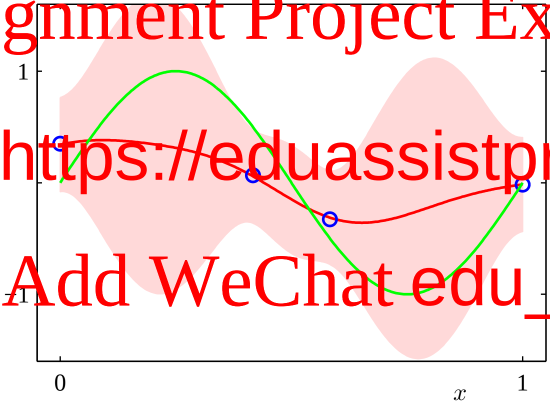
Example with artificial sinusoidal data from $\sin(2\pi x)$ (green) and added noise. Number of data points $N = 2$.



Mean of the predictive distribution (red) and regions of one standard deviation from mean (red shaded).

Predictive Distribution with Simplified Prior

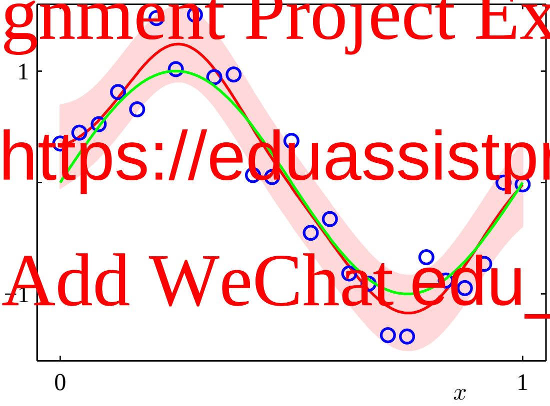
Example with artificial sinusoidal data from $\sin(2\pi x)$ (green) and added noise. Number of data points $N = 4$.



Mean of the predictive distribution (red) and regions of one standard deviation from mean (red shaded).

Predictive Distribution with Simplified Prior

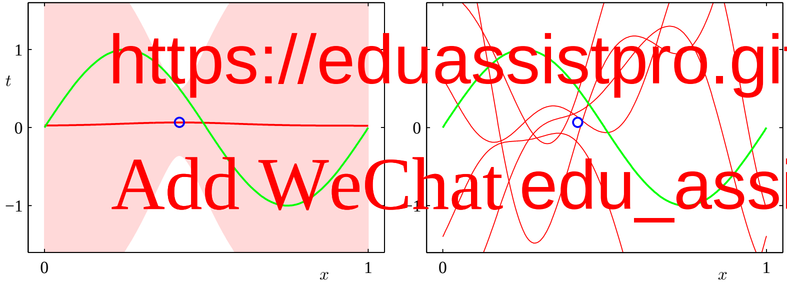
Example with artificial sinusoidal data from $\sin(2\pi x)$ (green) and added noise. Number of data points $N = 25$.



Mean of the predictive distribution (red) and regions of one standard deviation from mean (red shaded).

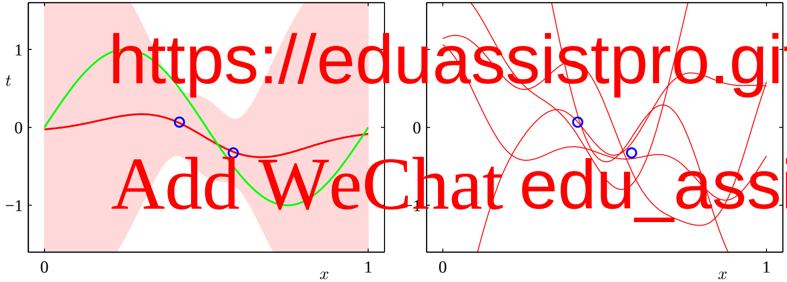
Predictive Distribution with Simplified Prior

Plots of the function $y(x, \mathbf{w})$ using samples from the posterior distribution over $\mathbf{w}$. Number of data points $N = 1$.



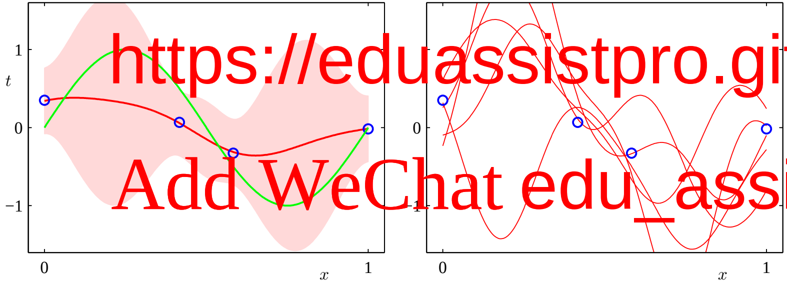
Predictive Distribution with Simplified Prior

Plots of the function $y(x, \mathbf{w})$ using samples from the posterior distribution over $\mathbf{w}$. Number of data points $N = 2$.



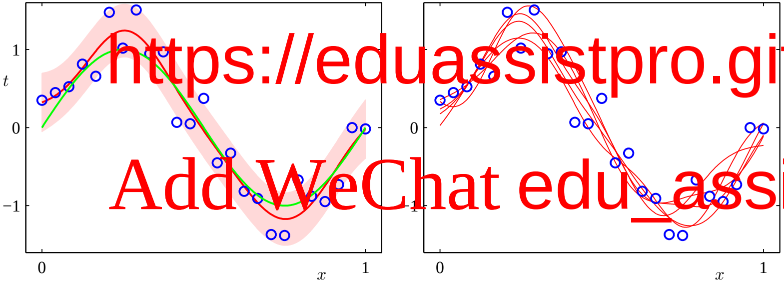
Predictive Distribution with Simplified Prior

Plots of the function $y(x, \mathbf{w})$ using samples from the posterior distribution over $\mathbf{w}$. Number of data points $N = 4$.



Predictive Distribution with Simplified Prior

Plots of the function $y(x, \mathbf{w})$ using samples from the posterior distribution over $\mathbf{w}$. Number of data points $N = 25$.



Limitations of Linear Basis Function Models

- Basis function $\phi_j(\mathbf{x})$ are fixed before the training data set is observed.
- Curse of dimensionality : Number of basis function grows rapi
- But t be e

  - dimension much smaller than D. Need al place basis functions only where data are (methods / Gaussian processes).
  - Target variables may only depend on a few sig directions within the data manifold. Need algorithms which can exploit this property (*e.g.* linear methods or shallow neural networks).

Curse of Dimensionality

- Linear Algebra allows us to operate in n-dimensional vector spaces using the intution from our 3-dimensional world as a vector space. No surprises as long as n is finite.
- If we add more structure to a vector space (e.g. inner pro
3-di

- Exa
the volume of the sphere in a D-dimensional space which lies between radius $r = 1$ and $r = 1 - \epsilon$
- Volume scales like r^D, therefore the formula of a sphere is $V_D(r) = K_D r^D$.

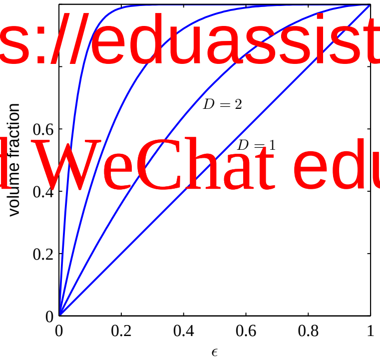
$$\frac{V_D(1) - V_D(1-\epsilon)}{V_D(1)} = 1 - (1-\epsilon)^D$$

Curse of Dimensionality

Ong & Walder & Webers & Xie
Data61 | CSIRO
ANU Computer Science

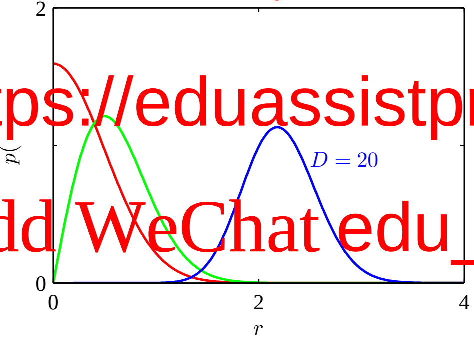
- Fraction of the volume of the sphere in a D-dimensional space which lies between radius $r = 1$ and $r = 1 - \epsilon$

$$\frac{V_D(1) - V_D(1 - \epsilon)}{V_D(1)} = 1 - (1 - \epsilon)^D$$



Curse of Dimensionality

- Probability density with respect to radius r of a Gaussian distribution for various values of the dimensionality D.



Curse of Dimensionality

- Probability density with respect to radius r of a Gaussian distribution for various values of the dimensionality D.
- Example: $D = 2$; assume $\mu = 0, \Sigma = I$

$$\mathcal{N}(x \mid 0, I) = \frac{1}{2\pi} \exp\left\{-\frac{1}{2} x^\top x\right\} = \frac{1}{2\pi} \exp\left\{-\frac{1}{2}(x_1^2 + x_2^2)\right\}$$

- Co

- Probability in the new coordinates

$$p(r, \phi \mid 0, I) = \mathcal{N}(r(x), \phi(x)) \mid J \mid$$

where $\mid J \mid = r$ is the determinant of the Jacobian for the given coordinate transformation.

$$p(r, \phi \mid 0, I) = \frac{1}{2\pi} r \exp\left\{-\frac{1}{2} r^2\right\}$$

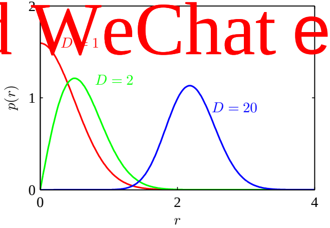
Curse of Dimensionality

- Probability density with respect to radius r of a Gaussian distribution for $D = 2$ (and $\mu = 0, \Sigma = I$)

$$p(r, \phi \,|\, 0, I) = \frac{1}{2\pi} r \exp\left\{-\frac{1}{2} r^2\right\}$$

- Inte

0



Summary: Linear Regression

- Basis functions
- Maximum likelihood with Gaussian noise
- Re
- Bia
- Co
- Bayesian linear regression
- Sequential update of the posterior
- Predictive distribution
- Curse of dimensionality