Imperial College London

May 13th

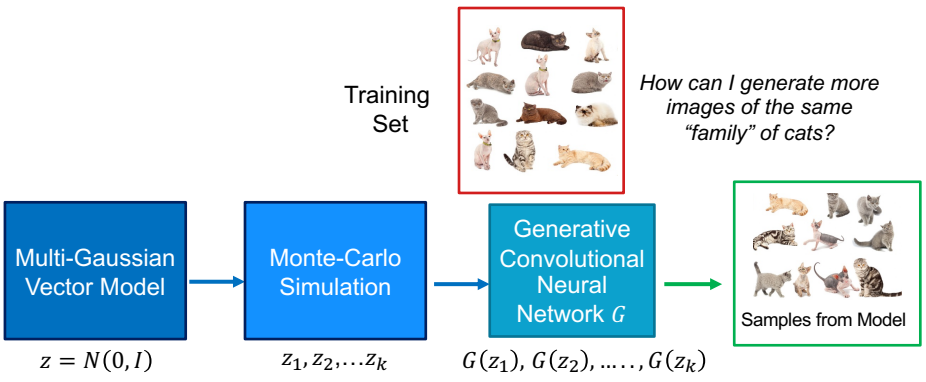
Probabilities for Deep Learning

Olivier Dubrule and Lukas Mosser

Imperial College London

Objectives of the Day

- Introduce Basic Probability Concepts Required to Understand Machine Learning, and in particular *Generative Networks*



Imperial College London

Probabilities for Deep Learning

1. Gaussian and Bernouilli distributions in one and n dimensions
2. Maximum Likelihood
3. Applying a Function to a Random Variable: Example of the Lognormal
4. Monte Carlo Simulation of a Gaussian variable or a Bernouilli variable
5. Comparing Probability Density Functions

Imperial College London

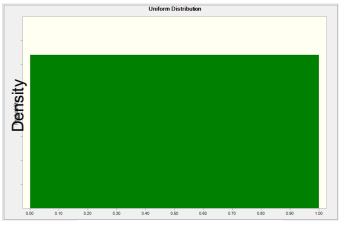
Defining a Random Variable

- A Random Variable X is a Variable that is associated with a non-deterministic process: we only know that it takes certain values with certain probabilities.
- Examples:
  - Outcome of the throw of a die
  - Outside temperature on January 1st, 2020 at noon in London
  - Before drilling a well, presence or absence of an oil field at a given location
  - Value of the Apple share next week same time
  - ….

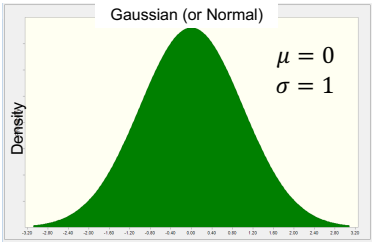
Imperial College London

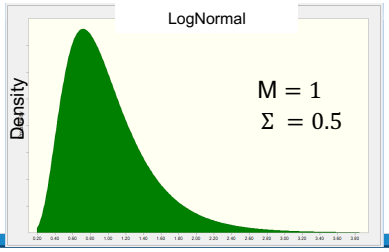
The PDF of a Continuous Random Variable

PDF = Probability Density Function



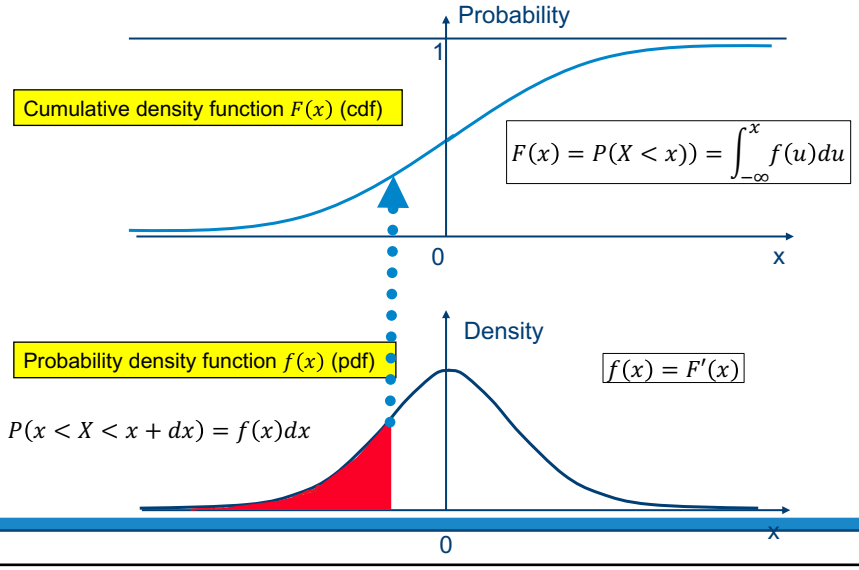
Uniform between 0 and 1





Imperial College London

Pdf and Cdf of a Random Variable X



Imperial College London

Properties of the pdf

$$\mathrm{P}(a < X < b) =) \int_a^b f(x)dx \qquad \int_{-\infty}^{+\infty} f(x)dx = 1$$

Mean $E(X) = \mu = \int_{-\infty}^{+\infty} xf(x)dx$

Variance $Var(X) = \sigma^2 = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x)dx$

$$= E[(X-\mu)^2] = E(X^2) - \mu^2$$

Imperial College London

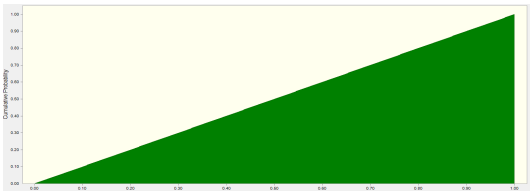
Exercise

Assume that $f(x)$ is a uniform pdf between 0 and 1.

Write its mathematical expression.

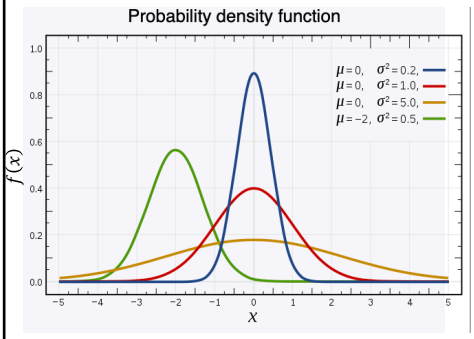
Calculate its mean and variance.

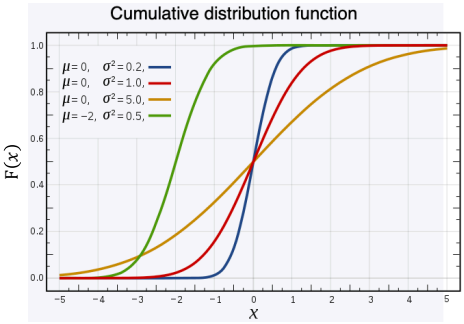
What is the mathematical expression of its cdf?



Imperial College London

The Normal (or Gaussian) Distribution





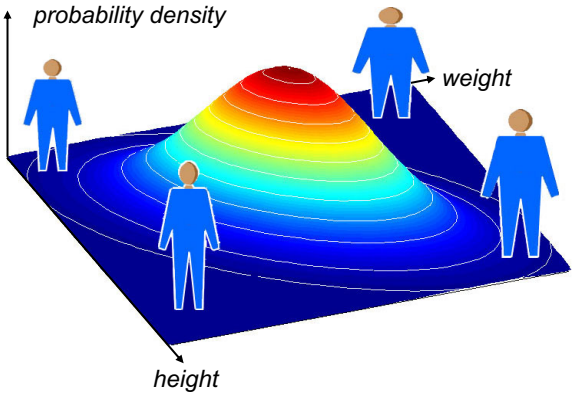
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$f(x)$ also written $N(x; \mu, \sigma^2)$

https://en.wikipedia.org/wiki/Normal_distribution

Imperial College London

Dealing with More Than One Dimension



Imperial College London

Covariance $Cov(X_1, X_2)$ of Two Random Variables X_1 and X_2

If X_1 has mean μ_1 and standard deviation σ_1
If X_2 has mean μ_2 and standard deviation σ_2

$$Cov(X_1, X_2) = E[(X_1 - \mu_1)(X_2 - \mu_2)]$$

The **correlation coefficient** ρ is defined as $\rho = \frac{Cov(X_1, X_2)}{\sigma_1 \sigma_2}$ or $Cov(X_1, X_2) = \rho\, \sigma_1 \sigma_2$

ρ has the following properties:

$$-1 \leq \rho \leq 1$$

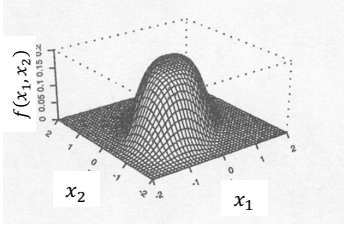
If $Cov(X_1, X_2) = 0$ or $\rho = 0$, then X_1 and X_2 are uncorrelated

If $\rho = -1$ or +1, there is a perfect linear relationship between X_1 and X_2

Imperial College London

Mathematical Expression of Bivariate Normal pdf

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{(x_1-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right)\right]$$



https://www.ilri.org/biometrics/Publication/Full%20Text/Linear_Mixed_Models/AppendixD.htm

Imperial College London

Exercise

Calculate the Bivariate Gaussian when

$$\mu_1 = \mu_2 = 0,\ \sigma_1 = \sigma_2 = 1$$

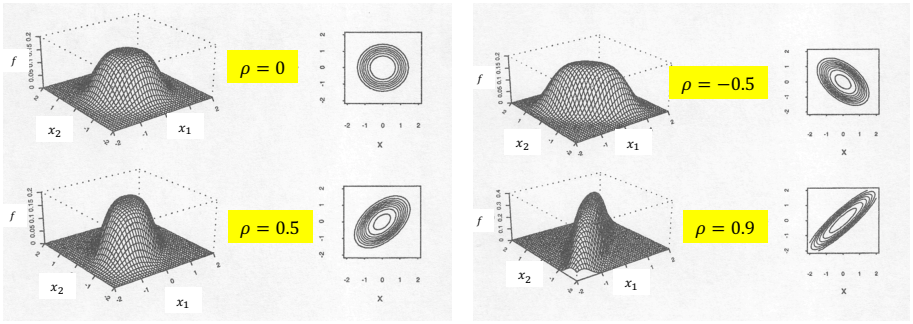
Then calculate it when $\rho = 0$, and show that it is the product of two Gaussians.

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{(x_1-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right)\right]$$

https://www.ilri.org/biometrics/Publication/Full%20Text/Linear_Mixed_Models/AppendixD.htm

Imperial College London

Examples of Bivariate Normal Distributions



Examples of Bivariate Normal Densities of (x_1, x_2) with $\mu_1 = \mu_2 = 0,\ \sigma_1 = \sigma_2 = 1$

https://www.ilri.org/biometrics/Publication/Full%20Text/Linear_Mixed_Models/AppendixD.htm

Imperial College London

Another Way to Write the Bivariate Normal Density

$$f(x_1,x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left(\frac{(x_1-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right)\right]$$

If Σ is the so-called Variance Covariance Matrix:

$$\Sigma = \begin{pmatrix} Var(X_1) & Cov(X_1,X_2) \\ Cov(X_1,X_2) & Var(X_2) \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

$$f(x_1,x_2) = \frac{1}{2\pi}\frac{1}{|\Sigma|}\exp\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right]$$

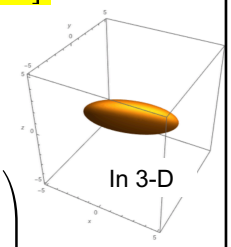
https://www.ilri.org/biometrics/Publication/Full%20Text/Linear_Mixed_Models/AppendixD.htm

Imperial College London

Multivariate Normal pdf of a Random Vector $(X_1, \ldots X_n)$

$$f(x) = (2\pi)^{-\frac{n}{2}}[\Sigma]^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right]$$

$\boldsymbol{x}$ is the $n\times1$ vector $\quad x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$



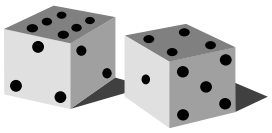
μ is the $n\times1$ expectation vector $\quad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} = \begin{pmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_n) \end{pmatrix}$

Σ is the $n\times n$ variance-covariance matrix $\quad \Sigma = \left(\left(Cov(X_i,X_j)\right)_{i,j=1,\ldots,n}\right)$

Imperial College London

Discrete Random Variable

- A DISCRETE random variable is a variable that can only take a finite number of values with a probability attached to each value
- Examples



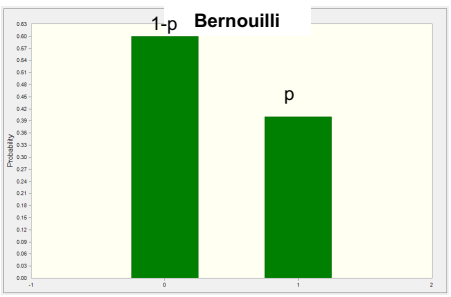
Throw of a die: X = 1, 2, 3, 4, 5, 6

Throw of a coin : Heads: X = -1 Tails: X = 1

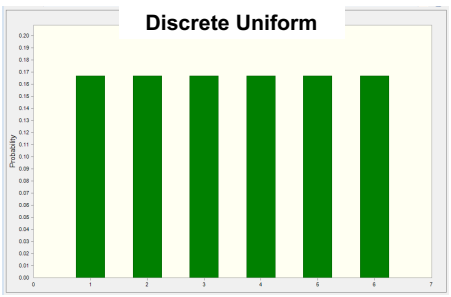
Bernouilli (or Indicator) variable: I=1 if sunshine tomorrow, I=0 if not

Imperial College London

The PDF of a Discrete Random Variable



Indicator Variable



Throw of a die

Imperial College London

Expectation of a Discrete Random variable

The expectation (or expected value) $E(X)$ of a discrete random variable X is the probability-weighted average of all possible outcome values. $E(X)$ is the theoretical mean of X.

Value of X	x_1	x_2	x_3	…	x_n
Probability	p_1	p_2	p_3	…	p_n

$$E(X) = p_1 x_1 + p_2 x_2 + p_3 x_3 + \cdots + p_n x_n$$

Exercise: What is the Expectation of the Throw of a Dice?

Imperial College London

A Bernouilli Discrete Random Variable

A Bernouilli random variable X only takes two possible values:
1 with probability p, and 0 with probability (1-p)

Value of X	1	0
Probability	p	1-p

What is the Expectation of X?

$E(X) = 1 \times p + 0 \times (1-p) = p$

A Bernouilli variable takes either the value 0 and 1, and we can write that the probability that it is equal to x is:

$$b(x) = p^x (1-p)^{1-x}$$

Exercise: What is the Variance of a Bernouilli Random Variable?

Imperial College London

Independent and Identically Distributed (IID)

*In probability theory, a sequence or collection of random variables is independent and identically distributed (**i.i.d.** or **iid** or **IID**) if each random variable has the same probability distribution as the others and they all are mutually independent.*

When treating m samples from a training or test dataset (for instance a set of images), it is assumed they are IID.

Imperial College London

Probabilities for Deep Learning

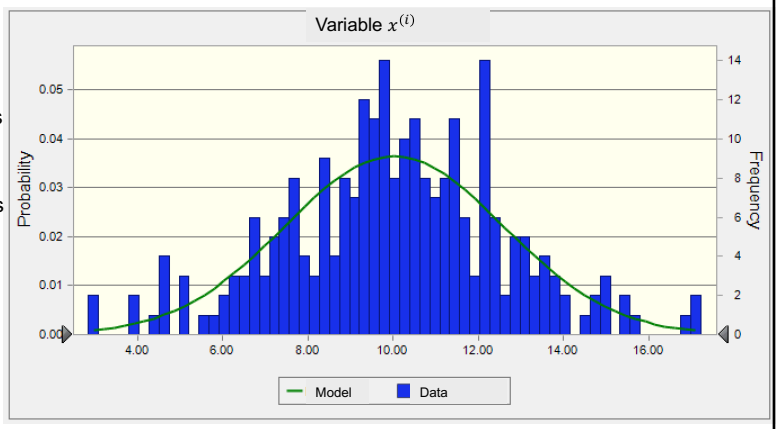
1. **Bernouilli and Gaussian distributions in one and n dimensions**
2. **Maximum Likelihood**
3. **Applying a Function to a Random Variable: Example of the Lognormal**
4. **Monte Carlo Simulation of a Gaussian variable or a Bernouilli variable**
5. **Comparing Probability Density Functions**

Imperial College London

Problem Adressed by the Maximum Likelihood Method

Number of Samples
$m = 250$

Number of Features
$n = 1$



Imperial College London

Problem Addressed by the Maximum Likelihood Method

Assume we have m *independent* (IID) data points $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$

Each data point is made of n features.

We want to fit a multivariate (for instance multivariate Gaussian) pdf $p_\theta(x)$ to these m samples.

Maximum likelihood consists of calculating the parameters θ such that the m samples maximize their likelihood.

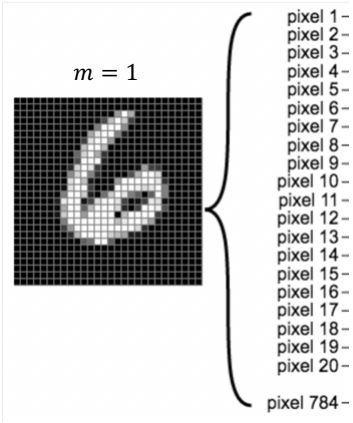
But what is the Likelihood?

Imperial College London

One Sample of MNIST

$n = 784$ (features)

$m = 1$



Imperial College London

Likelihood and Maximum Likelihood for a Dataset of Images

The pdf $p_\theta(x_1, x_2, \ldots, x_n)$ is parametrized by θ. If there is just one image $x^{(1)} = \left(x_1^{(1)}, x_2^{(1)}, \ldots, x_n^{(1)}\right)$ in the dataset, its likelihood is defined as:

$$Likelihood\ of\ image\ x^{(1)} = p_\theta\left(x^{(1)}\right) = p_\theta\left(x_1^{(1)}, x_2^{(1)}, \ldots, x_n^{(1)}\right)$$

The Maximum Likelihood estimate θ_{ML} of θ is calculated as

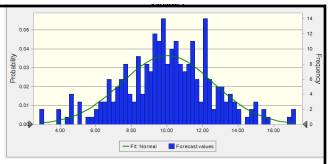
$$\theta_{ML} = argmax\left(p_\theta\left(x^{(1)}\right)\right) \quad \text{or} \quad \theta_{ML} = argmax\left(\log p_\theta\left(x^{(1)}\right)\right)$$

If there are m *IID* images $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$ in the dataset

$$\theta_{ML} = argmax\left(p_\theta\left(x^{(1)}\right)p_\theta\left(x^{(2)}\right)\ldots p_\theta\left(x^{(m)}\right)\right)$$

$$= argmax\left(\log\left(p_\theta\left(x^{(1)}\right)p_\theta\left(x^{(2)}\right)\ldots p_\theta\left(x^{(m)}\right)\right)\right) = \theta = argmax\left(\sum_{i=1}^{m}\log p_\theta\left(x^{(i)}\right)\right)$$

Imperial College London



Maximum Likelihood in the Gaussian Case

Goal : fit a Gaussian pdf to a dataset

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

Here the θ parameters are μ and σ

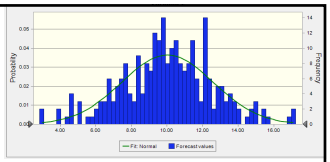
We have $p_\theta(x) = N(x;\mu,\sigma^2)$

Hence the Likelihood $\sum_{i=1}^{m}\left(\log p_\theta\left(x^{(i)}\right)\right) = -\frac{1}{2}\sum_{i=1}^{m}\left(\frac{x_i-\mu}{\sigma}\right)^2 - m\log\sigma - \frac{m}{2}\log 2\pi$

The Maximum Likelihood Estimate for a Gaussian leads to the L2 norm!

https://stats.stackexchange.com/questions/351549/maximum-likelihood-estimators-multivariate-gaussian
https://onlinecourses.science.psu.edu/stat414/node/191/

Imperial College London



Exercise

We have m IID values of real numbers $(x_i)_{i=1\ldots m}$

We want to calculate the parameters of a normal distribution $N(x;\mu,\sigma^2)$ that best fits these *m* values.

What is the maximum likelihood estimate μ_{ML} of μ?

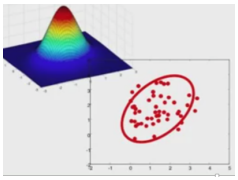
What is the maximum likelihood estimate σ^2_{ML} of σ^2?

$$\mu_{ML} = \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \sigma^2_{ML} = \frac{1}{m}\sum_{i=1}^{m}(x_i-\mu)^2$$

Imperial College London

Maximum Likelihood in the Multivariate Gaussian Case (1)

Idea : fit a multivariate Gaussian to a n-dimension dataset



$$f(x) = (2\pi)^{-\frac{n}{2}}[\Sigma]^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right]$$

Here the θ parameters are the vector μ and the matrix Σ ($|\Sigma|$ is its determinant)

With the previous notation we have $p_\theta(x) = N(x;\mu,\Sigma)$

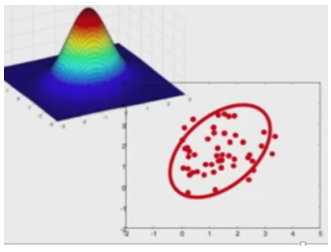
Hence

$$\sum_{i=1}^{m}\left(\log p_\theta\left(x^{(i)}\right)\right) = -\frac{nm}{2}\log 2\pi - \frac{m}{2}log|\Sigma| - \frac{1}{2}\sum_{i=1}^{m}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)$$

https://stats.stackexchange.com/questions/351549/maximum-likelihood-estimators-multivariate-gaussian

Imperial College London

Maximum Likelihood in the Multivariate Gaussian Case (2)



Maximum Likelihood Estimates

$$\mu_{ML} = \frac{1}{m}\sum_{i=1}^{m}x^{(i)}$$

$$\Sigma_{ML} = \frac{1}{m}\sum_{i=1}^{m}\left(x^{(i)}-\mu_{ML}\right)^T\left(x^{(i)}-\mu_{ML}\right)$$

https://stats.stackexchange.com/questions/351549/maximum-likelihood-estimators-multivariate-gaussian

Imperial College London

Maximum Likelihood for a Bernouilli Distribution

A Bernouilli distribution has just one parameter p

Suppose we have just one sample x_1 (which has the value 1 or zero)

Its likelihood is: $b(x) = p^{x_1}(1-p)^{1-x_1}$

Log-likelihood is: $x_1 \log p + (1-x_1)\log(1-p)$

If we have m samples x_i, their log-likelihood is:

$\sum_{i=1\ldots m}(x_i \log p + (1-x_i)\log(1-p))$ **(minus the cross-entropy!)**

The maximization leads, unsurpringly, to : $p_{ML} = \frac{1}{m}\sum_{i=1\ldots m} x_i$

Imperial College London

Probabilities for Deep Learning

1. **Bernouilli and Gaussian distributions in one and n dimensions**
2. **Maximum Likelihood**
3. **Applying a Function to a Random Variable: Example of the Lognormal**
4. **Monte Carlo Simulation of a Gaussian variable or a Bernouilli variable**
5. **Comparing Probability Density Functions**

Imperial College London

Pdf of the Function Y of a Random Variable X: $Y = h(X)$

X has pdf $f(x)$ and cdf $F(x)$
Y has pdf $g(y)$ and cdf $G(y)$

We have: $G(y) = P(Y < y) = P(h(X) < y) = P\left(X < h^{-1}(y)\right) = F\left(h^{-1}(y)\right)$

Hence: $g(y) = \mathrm{G}'(y)$

Exercise: What is e^X if X is Gaussian? $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{x-\mu}{\sigma}\right)^2}$ and $Y = h(X) = e^X$

$G(y) = P(Y < y) = P(e^X < y) = P(X < Logy)$

$$g(y) = G'(Logy) = \frac{1}{y}\frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{\mathrm{Log}\, y-\mu}{\sigma}\right)^2}$$

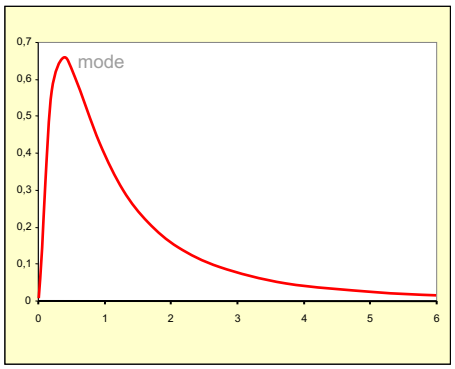
Lognormal pdf!

See Exercise 4

Imperial College London

X is Lognormally Distributed if $\log X$ is Normally Distributed

Distribution of $\mathrm{Y} = e^X$ when X is normal



	$\log Y$	Y
mean	μ	M
variance	σ^2	Σ^2

$$f(y) = \frac{1}{y}\frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{\log y-m}{\sigma}\right)^2}$$

$$Mean\ \ M = e^{m+\frac{\sigma^2}{2}}$$

Imperial College London

Generalizing the Change of Variable to Multidimensional Space

Suppose we have an n-dimensional random vector $X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}$ with a pdf

$$p(x) = p(x_1, x_2, \quad \dots \quad , x_n)$$

We apply to it the function g, where g is a bijective function from R^n to R^n.

We write : $\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = g \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$. Consider the Jacobian matrix $J = \left(\frac{\partial y_i}{\partial x_j}\right)$.

We have the relationship:

$$\boldsymbol{p(y_1, \quad \dots \quad , y_n) = p(x_1 \quad \dots \quad , x_n) \left|\frac{\partial y_i}{\partial x_j}\right|^{-1}}$$

Imperial College London

Exercise

Suppose that the two random variables X_1 and X_2 are each $N(x; 0,1)$ and independent from each other.

Consider the transformation: $\begin{aligned} Y_1 &= 4\,X_1 + 3\,X_2 \\ Y_2 &= X_1 + X_2 \end{aligned}$

What is the pdf of the bivariate random variable (Y_1, Y_2)?

$$Jacobian\, J = \begin{pmatrix} 4 & 3 \\ 1 & 1 \end{pmatrix} \qquad |J| = 1$$

$$g(y_1, y_2) = \mathrm{f}(x_1, x_2) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x_1^2}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x_2^2}{2}} = \frac{1}{2\pi} e^{-\frac{x_1^2 + x_2^2}{2}} = \frac{1}{2\pi} e^{-25\left(\frac{y_1^2}{25} + \frac{y_2^2}{2} - \frac{7}{25} y_1 y_2\right)}$$

Imperial College London

Probabilities for Deep Learning

1. Bernouilli and Gaussian distributions in one and n dimensions
2. Maximum Likelihood
3. Applying a Function to a Random Variable: Example of the Lognormal
4. Monte Carlo Simulation of a Gaussian variable or a Bernouilli variable
5. Comparing Probability Density Functions

Imperial College London

Initialization of Neural Network Parameters

From AlexNet paper

5 Details of learning

We trained our models using stochastic gradient descent with a batch size of 128 examples, momentum of 0.9, and weight decay of 0.0005. We found that this small amount of weight decay was important for the model to learn. In other words, weight decay here is not merely a regularizer: it reduces the model's training error. The update rule for weight w was

$$v_{i+1} := 0.9 \cdot v_i - 0.0005 \cdot \epsilon \cdot w_i - \epsilon \cdot \left\langle \frac{\partial L}{\partial w}|_{w_i} \right\rangle_{D_i}$$

$$w_{i+1} := w_i + v_{i+1}$$

where i is the iteration index, v is the momentum variable, ϵ is the learning rate, and $\left\langle \frac{\partial L}{\partial w}|_{w_i} \right\rangle_{D_i}$ is the average over the ith batch D_i of the derivative of the objective with respect to w, evaluated at w_i.

We initialized the weights in each layer from a zero-mean Gaussian distribution with standard deviation 0.01. We initialized the neuron biases in the second, fourth, and fifth convolutional layers, as well as in the fully-connected hidden layers, with the constant 1. This initialization accelerates the early stages of learning by providing the ReLUs with positive inputs. We initialized the neuron biases in the remaining layers with the constant 0.

We used an equal learning rate for all layers, which we adjusted manually throughout training. The heuristic which we followed was to divide the learning rate by 10 when the validation error rate stopped improving with the current learning rate. The learning rate was initialized at 0.01 and

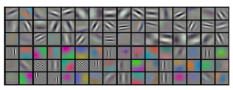


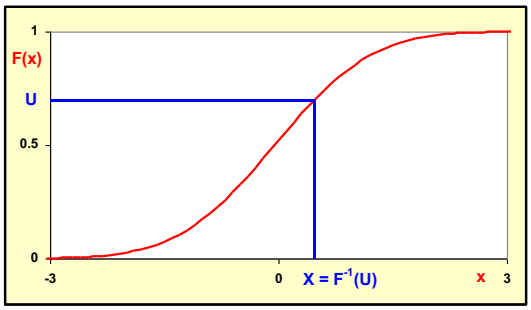
Figure 3: 96 convolutional kernels of size $11 \times 11 \times 3$ learned by the first convolutional layer on the $224 \times 224 \times 3$ input images. The top 48 kernels were learned on GPU 1 while the bottom 48 kernels were learned on GPU 2. See Section 6.1 for details.

6

Imperial College London

Simulation of a Continuous Random Variable with cdf F(x)

Draw a uniform number U between zero and one



$$P(X < x) = P\left(F^{-1}(U)\right) < x = P\left(U < F(x)\right) = F(x)$$

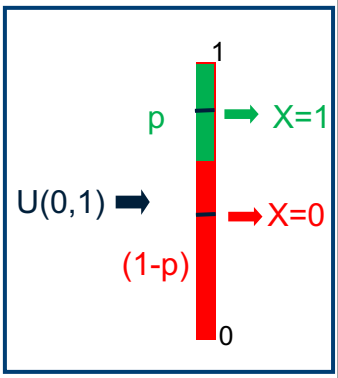
Find the value X associated with U using the formula $X=F^{-1}(U)$

Imperial College London

Simulation of a Bernouilli Random Variable

Assumption: X takes values 1 or 0 with probabilities p and $(1-p)$

Method: On the [0,1] interval define segments of lengths p and $(1-p)$. Simulate a uniform [0,1] and select the value corresponding to the segment in which the random point falls.



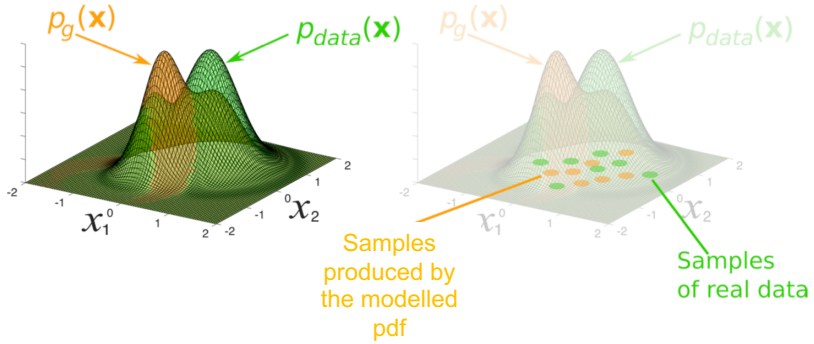
Imperial College London

Probabilities for Deep Learning

1. Bernouilli and Gaussian distributions in one and n dimensions
2. Maximum Likelihood
3. Applying a Function to a Random Variable: Example of the Lognormal
4. Monte Carlo Simulation of a Gaussian variable or a Bernouilli variable
5. Comparing Probability Density Functions

Imperial College London

Need to Compare pdfs when Modelled pdfs are Fitted to Data



Imperial College London

Compare two pdfs: the Kullback-Leibler (KL) Divergence

If the distributions $p(x)$ and $q(x)$ are continuous:

$$D_{KL}(p\|q) = \int_{-\infty}^{+\infty} p(x)\log p(x)\,dx - \int_{-\infty}^{+\infty} p(x)\log q(x)\,dx$$

Two Fundamental Properties

- $D_{KL}(p\|q)$ is positive, and 0 if $p(x)$ and $q(x)$ identical
- Asymmetry: $D_{KL}(p\|q) \neq D_{KL}(q\|p)$

Imperial College London

KL Divergence between Gaussians $N(x, \mu_1, \sigma_1^2)$ and $N(x, \mu_2, \sigma_2^2)$

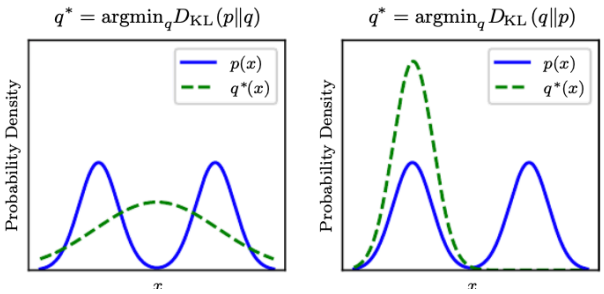
$$KL(f,g) = -\frac{1}{2} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} + \log\frac{\sigma_2}{\sigma_1}$$

(see Exercise 3 for calculation of KL Divergence for Gaussians and Multi-Gaussians)

From Deep Learning, by Goodfellow et al, 2016

Imperial College London

KL Divergence is Asymmetric



Depending on the choice of the KL Divergence used, the pdf $q^*(x)$ fitted to the pdf $p(x)$ will be different!

From Deep Learning, by Goodfellow et al, 2016

Imperial College London

KL Divergence versus Maximum Likelihood

If there are m *IID* images or samples $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$, the Maximum Likelihood estimate of the parameter θ associated with the modelled distribution p_θ is

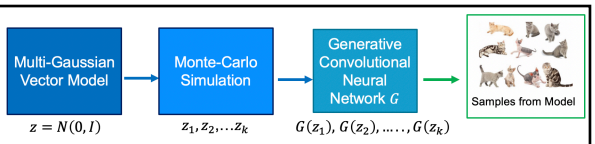
$$\theta = argmax\left(\frac{1}{m}\sum_{i=1}^{m}\log p_\theta\left(x^{(i)}\right)\right)$$

If we compare this expression to the expression of $D_{KL}(p\|q)$, we see that

$$argmax\left(\tfrac{1}{m}\textstyle\sum_{i=1}^{m}\log p_\theta\left(x^{(i)}\right)\right) = -argmin\left(D_{KL}\left(p_{data}\|p_\theta\right)\right)$$

Finding the parameters of the pdf which maximize the likelihood is equivalent to finding the parameters which minimize the KL Divergence.

Imperial College London



Conclusion

- The basic pdfs used in Deep Learning are Bernouilli and (Multivariate) Gaussian.
- There are formulas to calculate the pdf of the function of another random variable whose pdf is known. These formulas are used for Monte-Carlo simulation.
- The Kullback-Leibner (KL) Divergence is used to calculate the similarity between two pdfs. Minimizing it allows one pdf to be adjusted to another.
- Maximum Likelihood is a key approach in Deep Learning and applying it to Bernouilli and (Multivariate) Gaussian random variables leads respectively to the minimization of the cross-entropy for classification or the L2 norm for regressions.