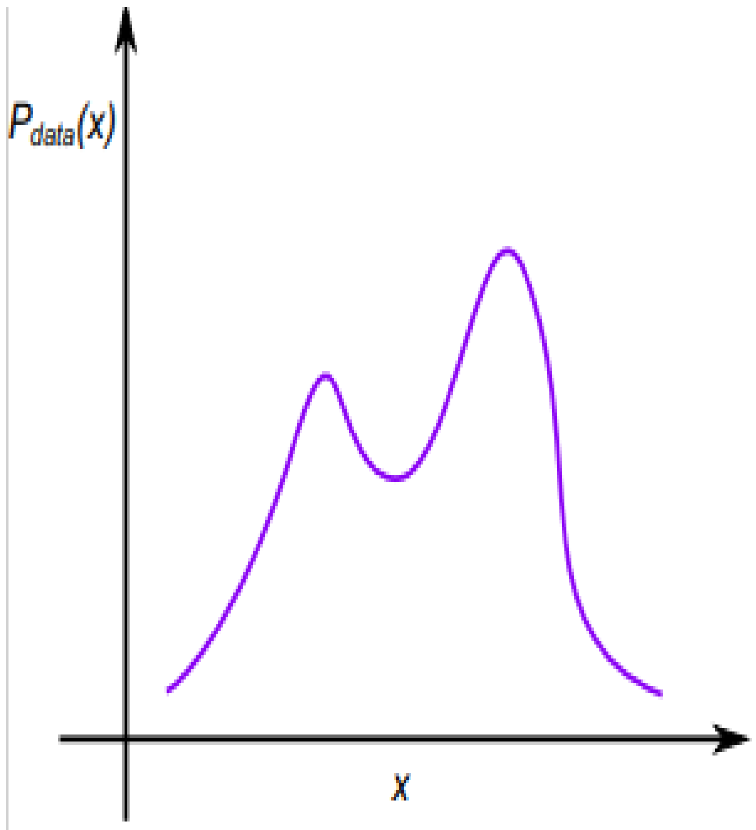
Probabilistic Deep Learning

Mustafa Hajij

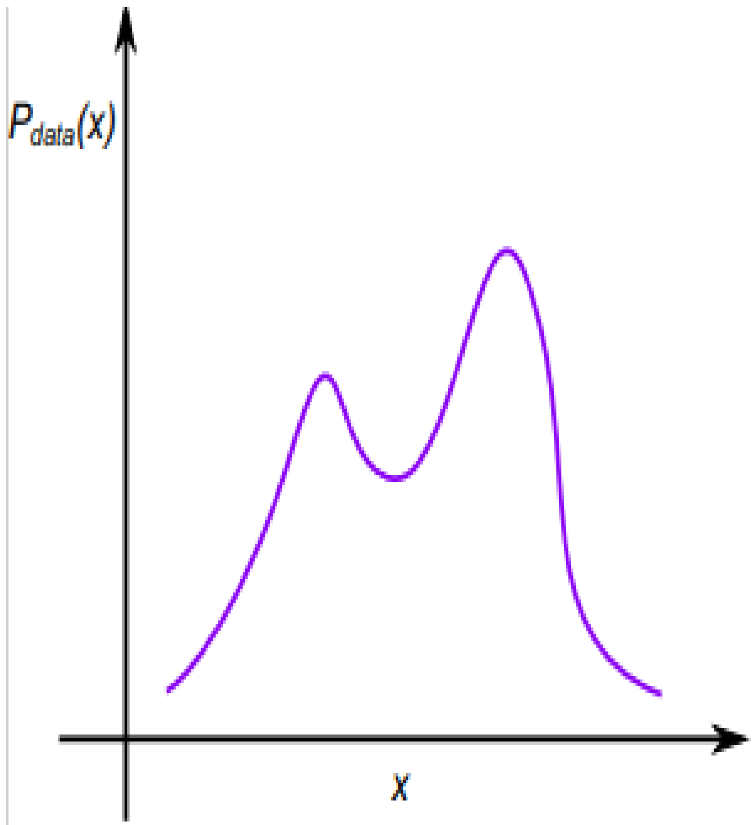
Introduction and motivation



$P_{data}(x)$ is an unknow probability distribution

Think about p_data as distrubition of natural images

Introduction and motivation



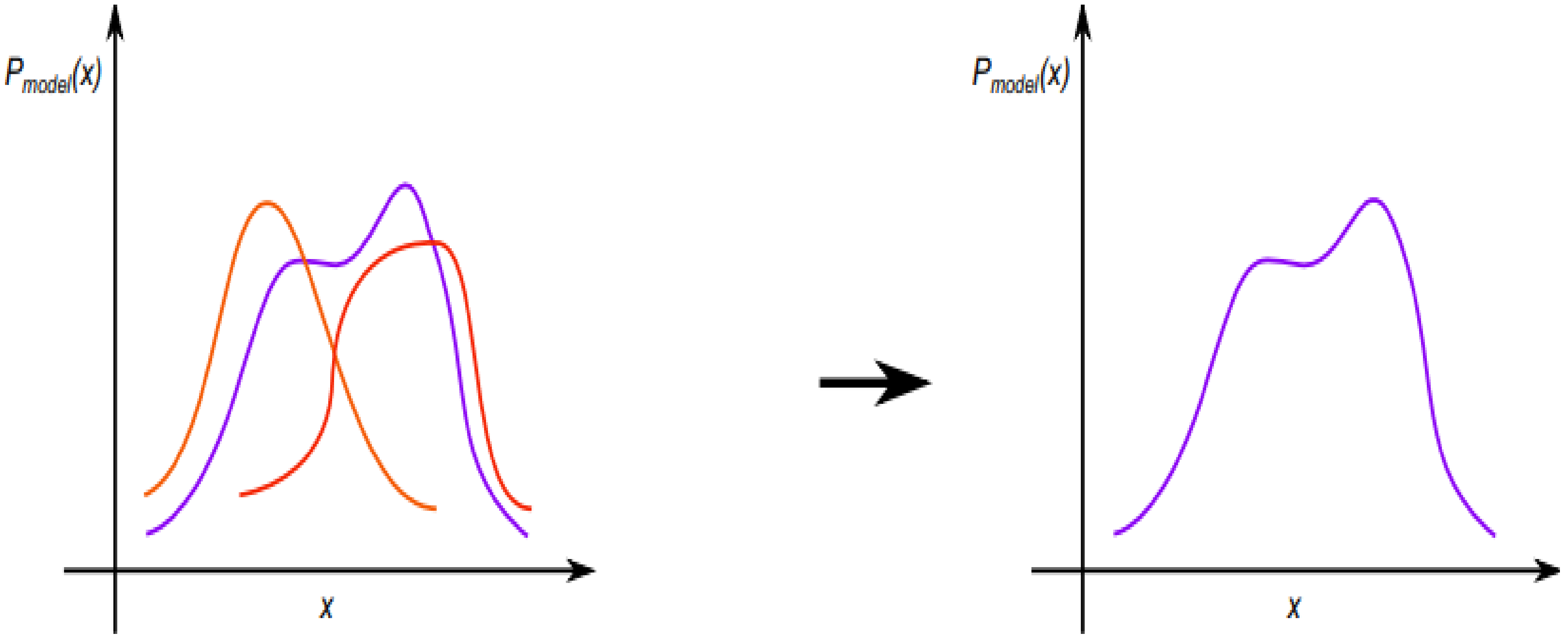
$P_{data}(x)$ is an unknow probability distribution

We do not have access to the mathematical formulations of that distribution but we have a cameras that can sample from it

Obtaining training data

Think about p_data as distrubition of natural images

Introduction and motivation



A machine learning algorithm searches the hypothesis space to find the right model.

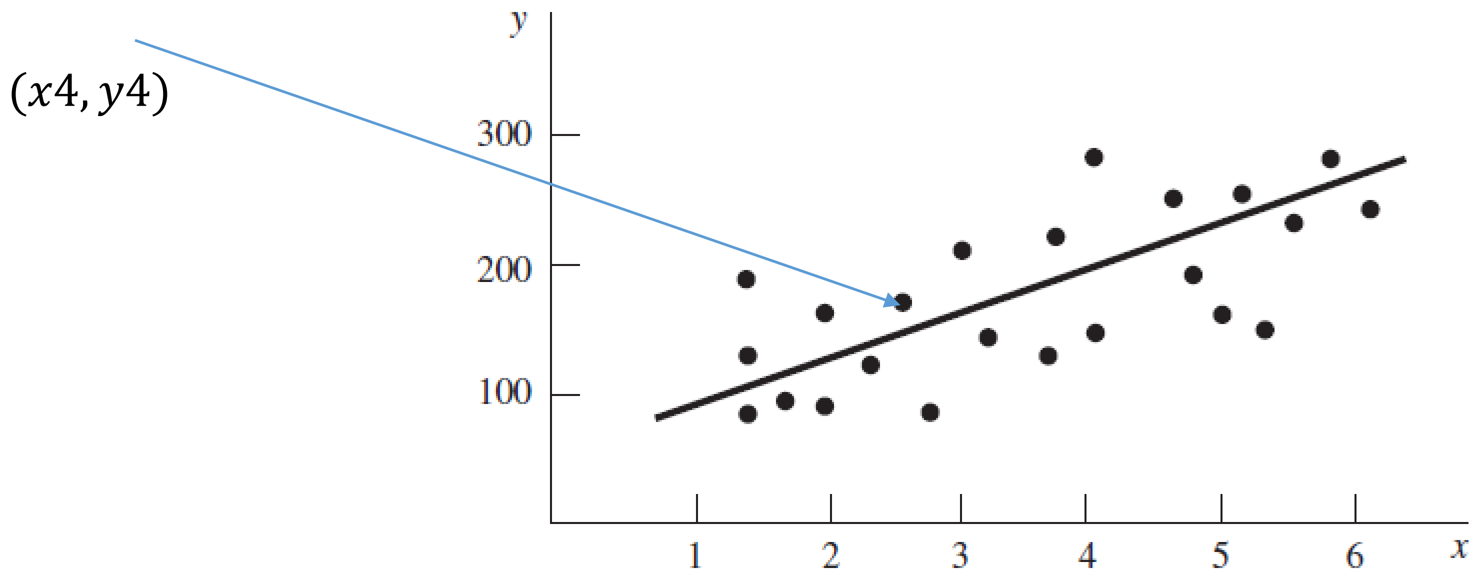
$P_{model}(x)$ models certain aspects about the original distribution $P_{data}(x)$.

Introduction and motivation

Suppose that you are given a collection of point $\{xi, yi\}_{\mathrm{i=1}}^{\mathrm{n}}$. We think of x_i as an independent variable and y_i as a dependent variable .

We are seeking to model the functional relationalship g between $x_i's$ and $y_i's$. In other words, want to find the function g such that

$$y = g\ (x;\ \boldsymbol{\beta})$$



Introduction and motivation

Suppose that you are given a collection of point
$\{xi, yi\}_{i=1}^{n}$. We think of x_i as an independent variable and y_i as a dependent variable .

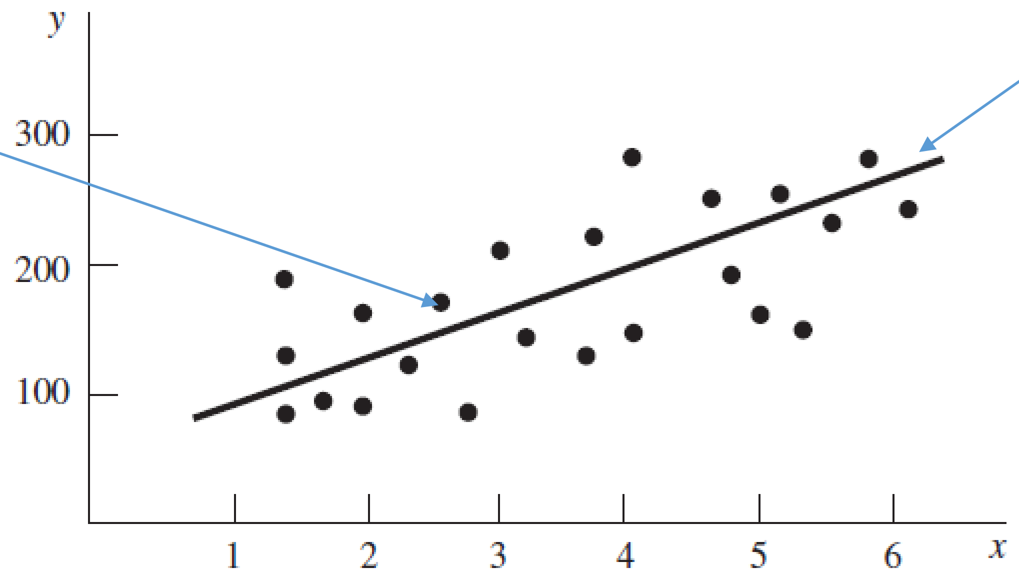
We are seeking to model the functional relationalship g between $x_i's$ and $y_i's$. In other words, want to find the function g such that

y = g (x; **β**)

In simple regression The function g is a line β 0 + β 1*x*

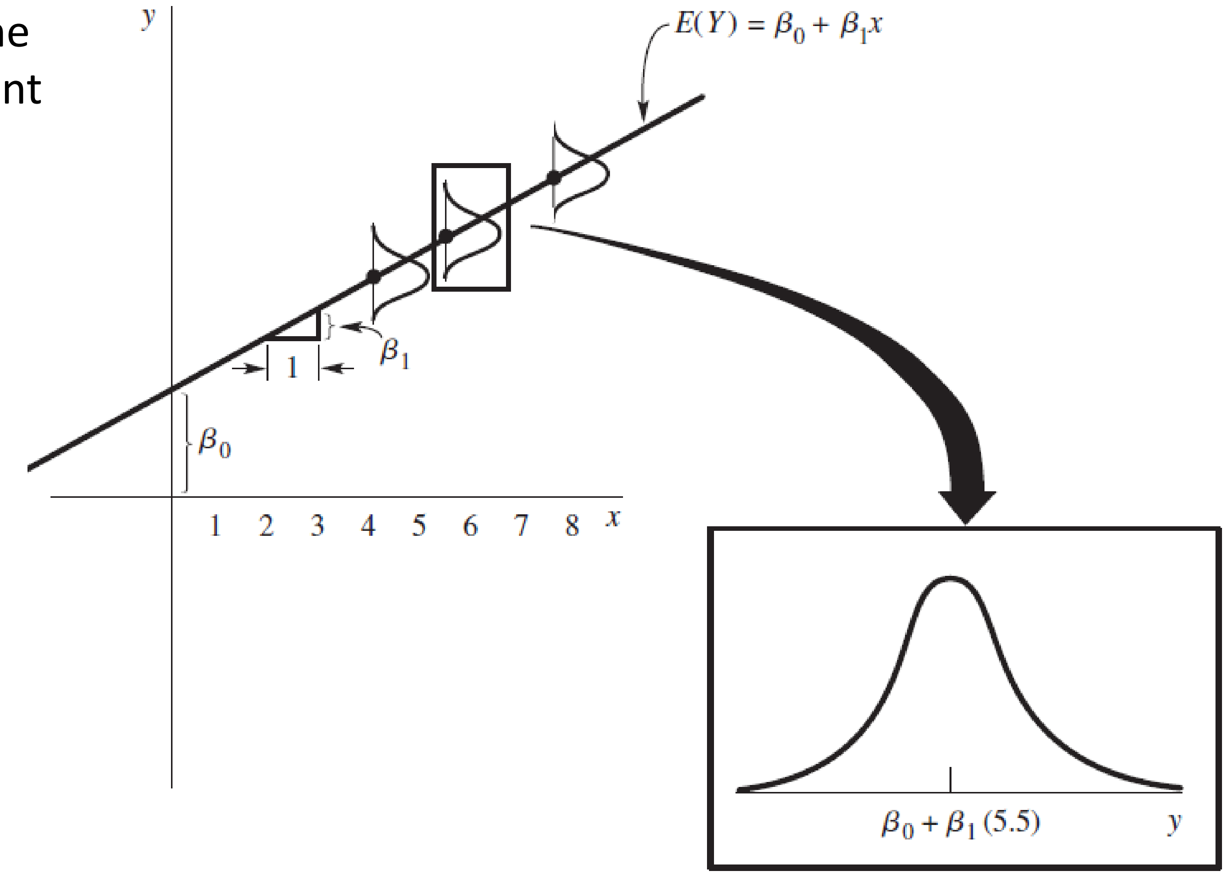
β= (β 0 , β 1)

$(x4, y4)$



Introduction and motivation

Key change of perspective : *In fact a better way to look at what we did* is that we modeled the distribution of y conditioned above every x point

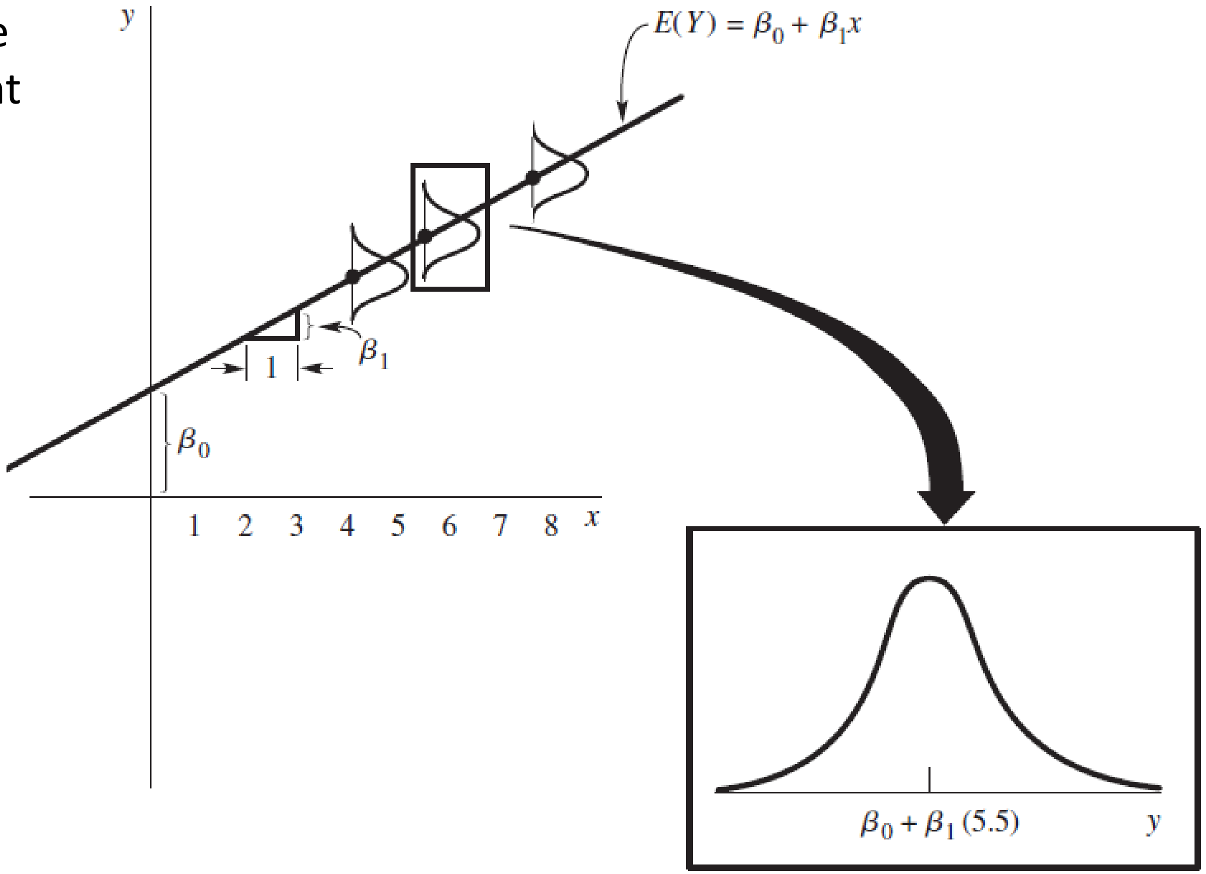


the model we studied

$$Y_{x_i} \sim N(\mu_{x_i} = \beta 0 + \beta 1 x_i \quad , \sigma^2_{x_i} = \sigma^2)$$

Introduction and motivation

Key change of perspective : $In\ fact\ a\ better\ way\ to\ look\ at\ what\ we\ did$ is that we modeled the distribution of y conditioned above every x point



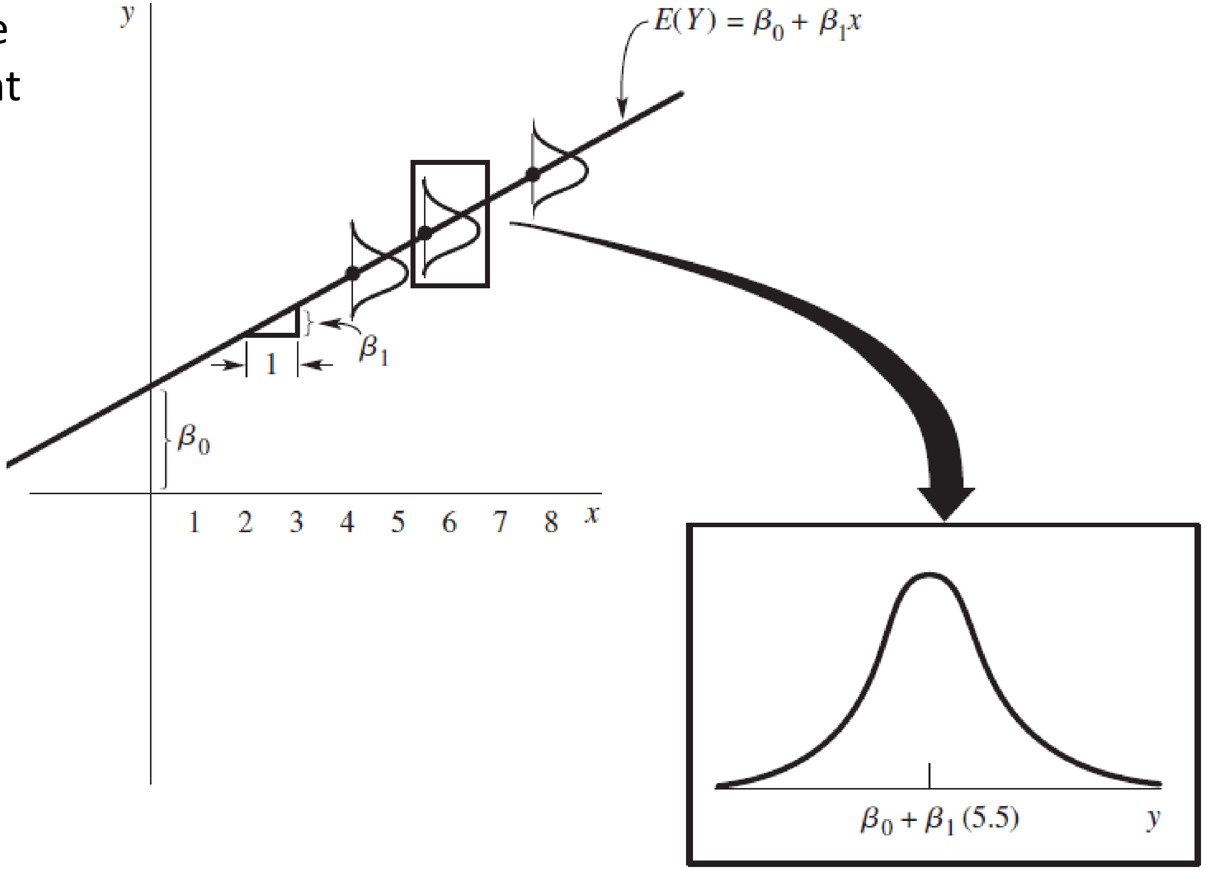
The way you look at it above every point x_i the random variable Y_{x_i} is a normal distribution With mean $\mu_{x_i} = \beta\, 0 + \beta\, 1 x_i$ and standard deviation σ^2

$the\ model\ we\ studied$

$$Y_{x_i} \sim N(\mu_{x_i} = \beta\, 0 + \beta\, 1 x_i \quad , \sigma^2_{x_i} = \sigma^2)$$

Introduction and motivation

Key change of perspective : *In fact a better way to look at what we did* is that we modeled the distribution of y conditioned above every x point



The way you look at it above every point x_i the random variable Y_{x_i} is a normal distribution With mean $\mu_{x_i} = \beta 0 + \beta 1 x_i$ and standard deviation σ^2

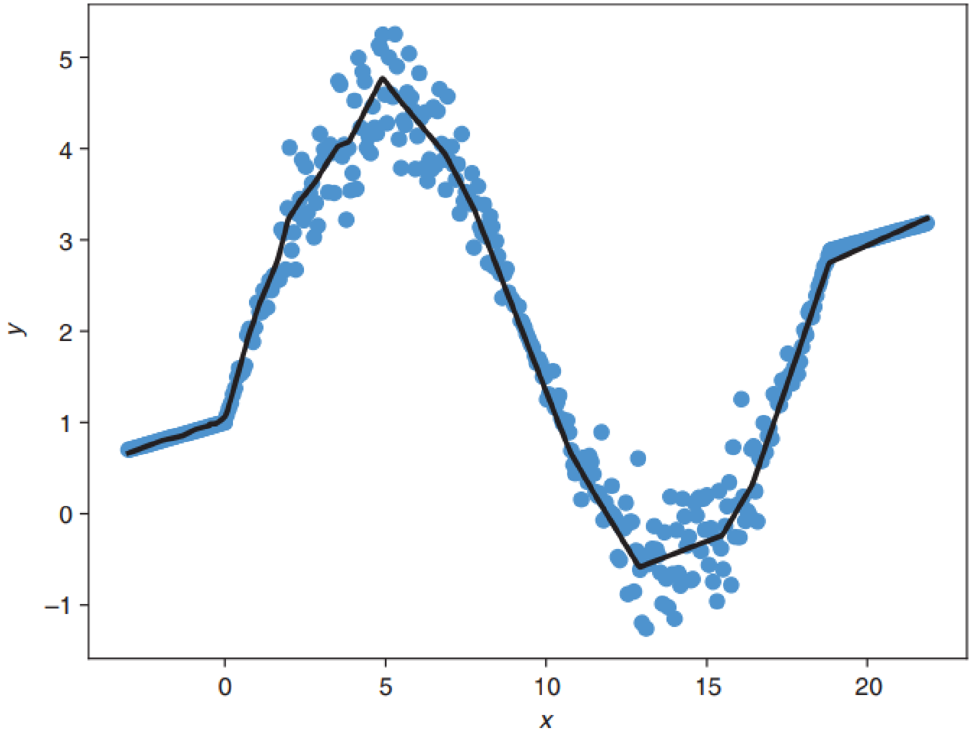
Since σ^2 is constant, in a typical regression problem We just focus on the mean of Y_x

the model we studied

$$Y_{x_i} \sim N(\mu_{x_i} = \beta 0 + \beta 1 x_i \quad , \sigma^2_{x_i} = \sigma^2)$$

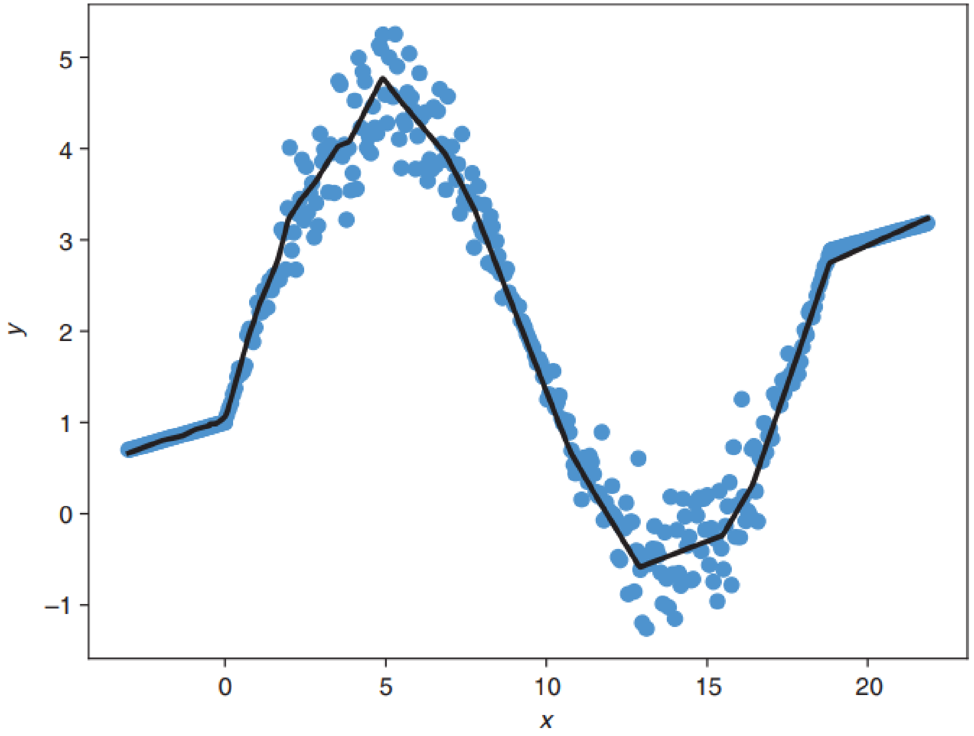
Introduction and motivation

But what if the variance $\sigma_{x_i}^2$ is depends on x_i ?



Introduction and motivation

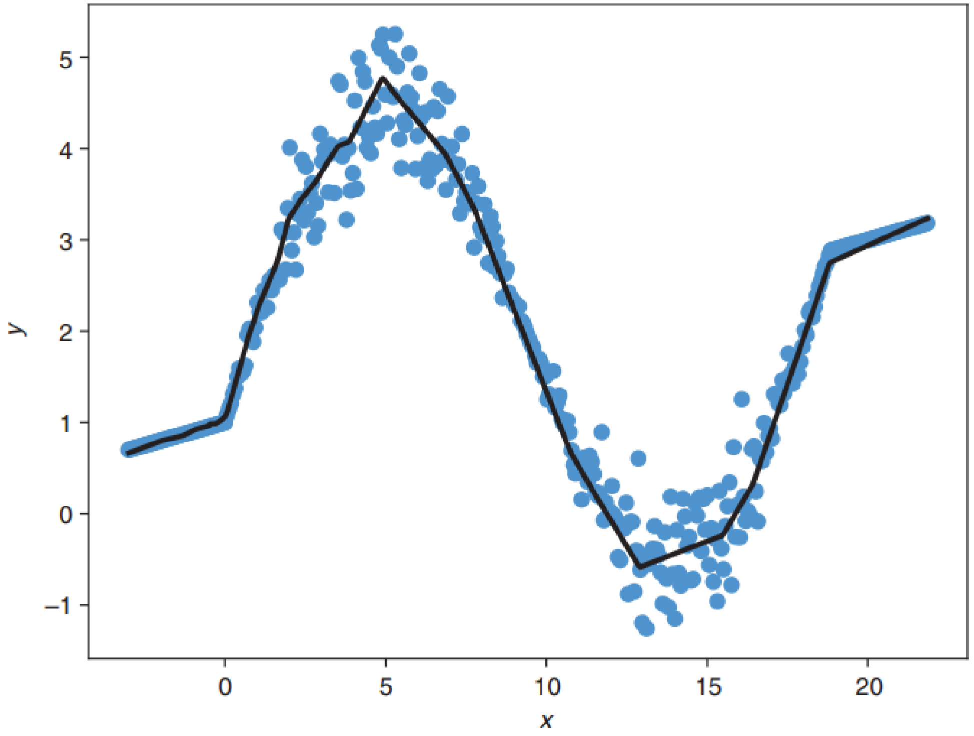
But what if the variance $\sigma^2_{x_i}$ is depends on x_i ?



$\sigma^2_{x_i}$ depends on x_i and it changes as we change it!

Introduction and motivation

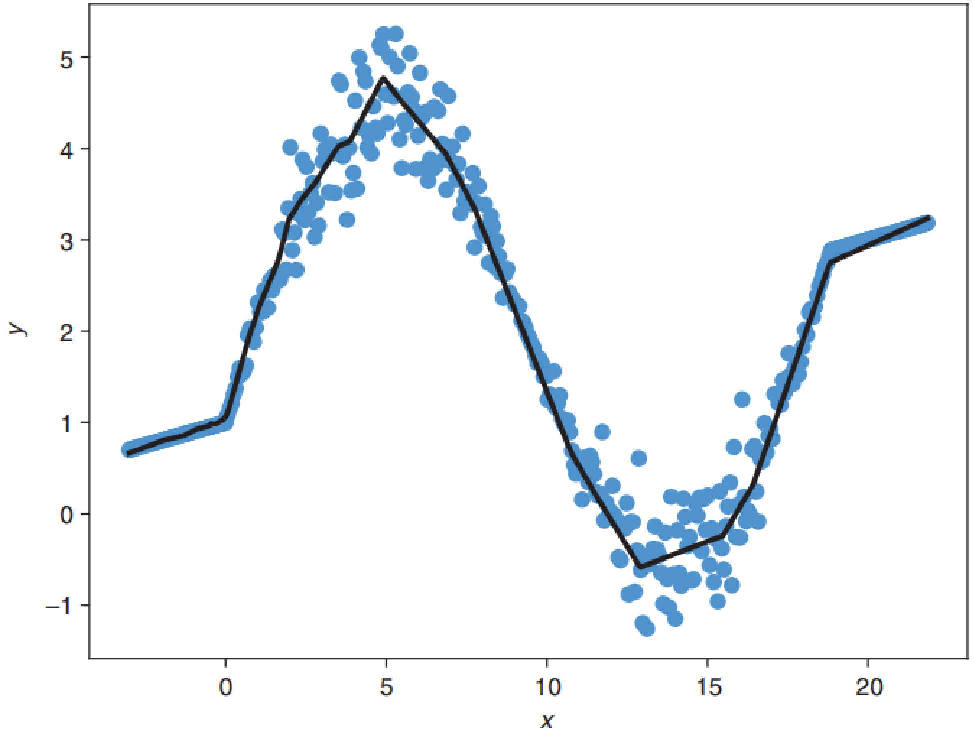
But what if the variance $\sigma^2_{x_i}$ is depends on x_i ?



$$More\ generally\ what\ if$$
$$Y_{x_i} \sim some\ unknown\ distribution?$$

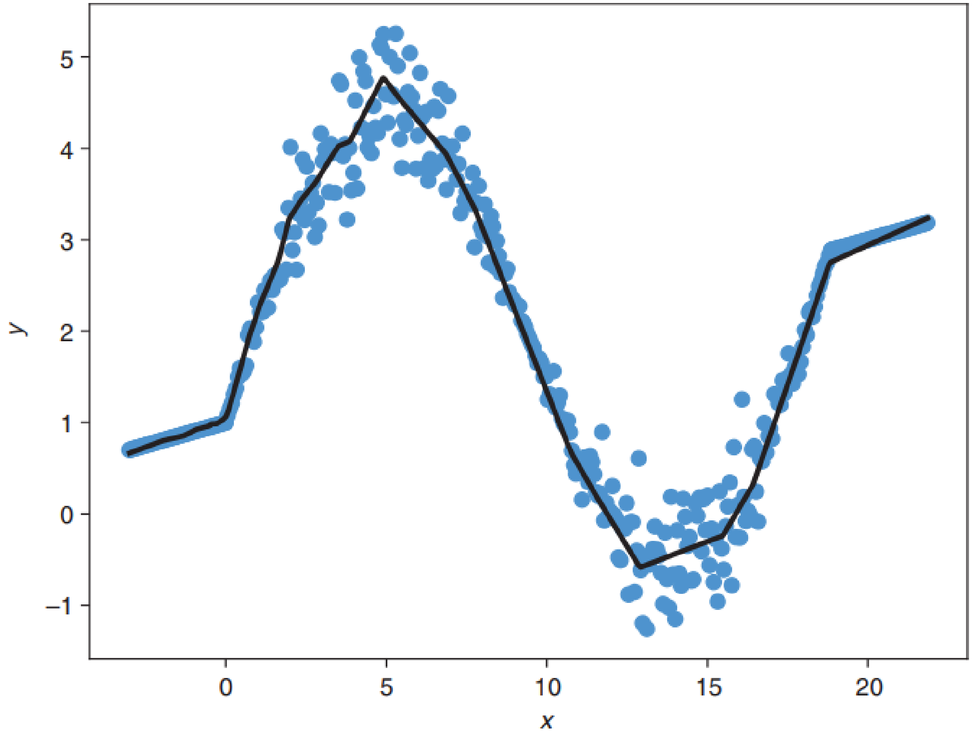
Introduction and motivation

Lets examine this case:



Introduction and motivation

Lets examine this case:



$Lets\ assume\ that Y_{x_i}$ is still normal but this time lets assume both mean and variance of Y_{x_i} are general functions :

$$Y_{x_i} \sim \mathrm{N}(\mu_{x_i} = \text{f1 } (x;\boldsymbol{\beta}), \sigma^2_{x_i} = \text{f2 } (x;\boldsymbol{\beta}))$$

Here

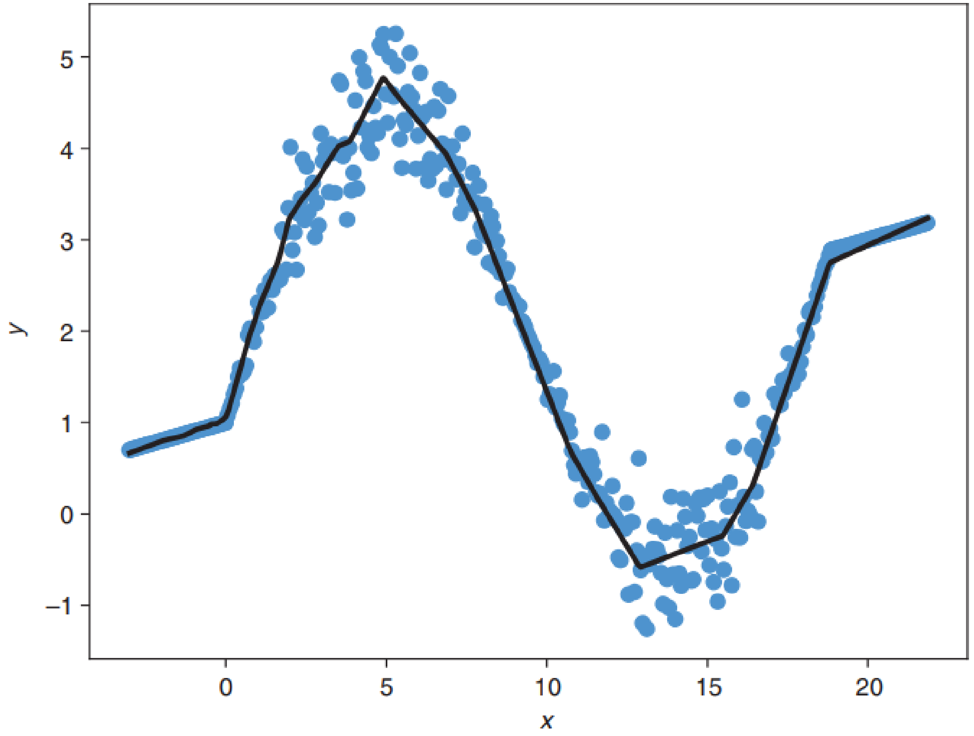
f1 (x;**β**)

f2 (x;**β**)

Are some non-linear functions!

Introduction and motivation

Lets examine this case:



Lets assume that Y_{x_i} is still normal but this time lets assume both mean and variance of Y_{x_i} are general functions :

$$Y_{x_i} \sim \mathrm{N}(\mu_{x_i} = \text{f1 }(x;\boldsymbol{\beta}), \sigma^2_{x_i} = \text{f2 }(x;\boldsymbol{\beta}))$$

Here

f1 (x;**β**)

f2 (x;**β**)

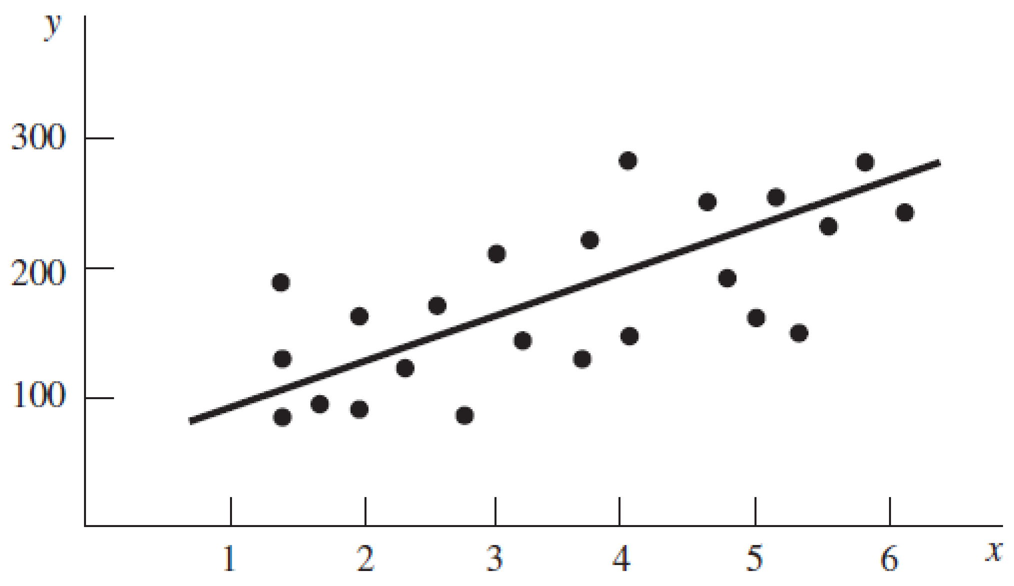
Are some non-linear functions!

Question : How can we convert the above math to a DL model?

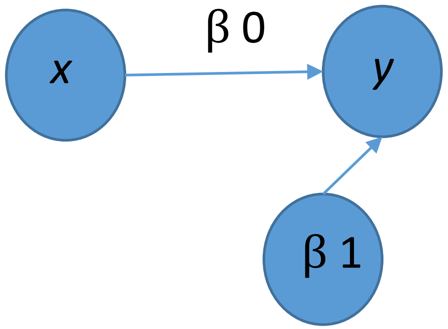
Introduction and motivation

In simple regression The function g is a line $\beta_0 + \beta_1 x$

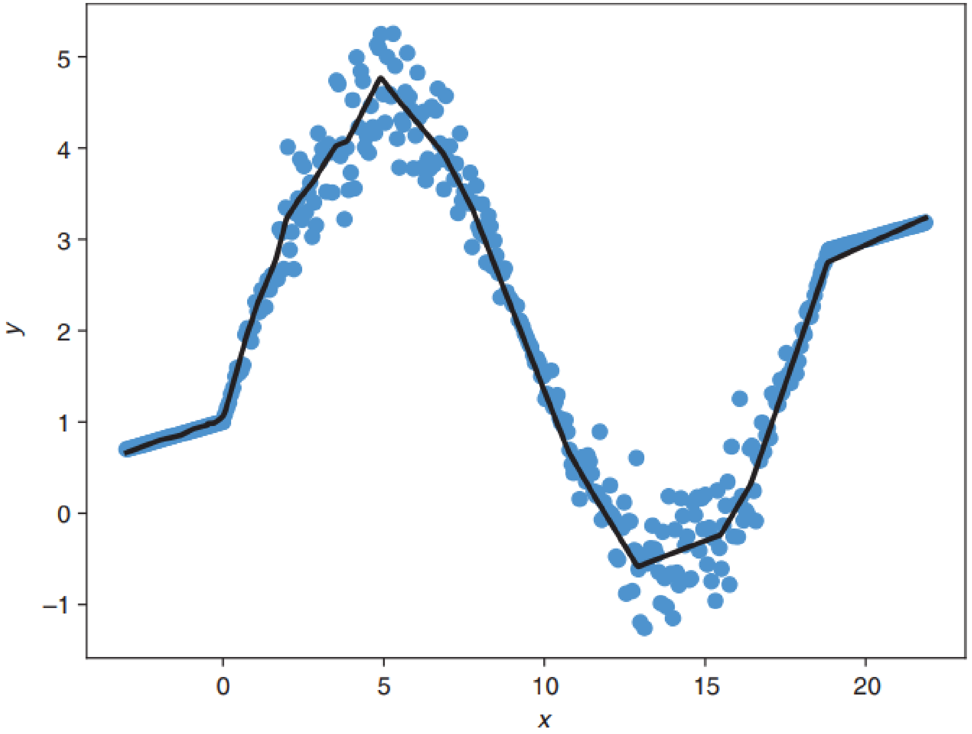
$\boldsymbol{\beta} = (\beta_0, \beta_1)$



Lets examine the simple case : the above regression as a neural network model



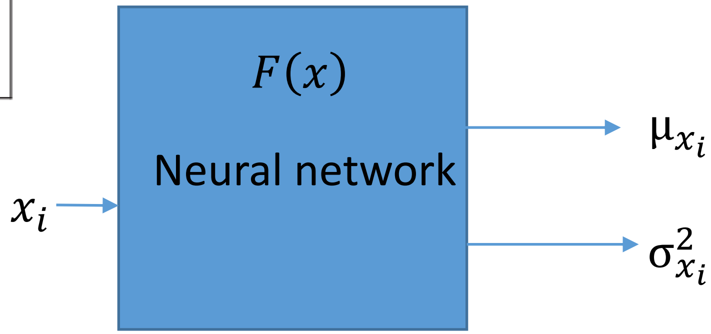
Introduction and motivation



$$Y_{x_i} \sim N(\mu_{x_i} = \text{f1 }(x;\boldsymbol{\beta}), \sigma^2_{x_i} = \text{f2 }(x;\boldsymbol{\beta}))$$

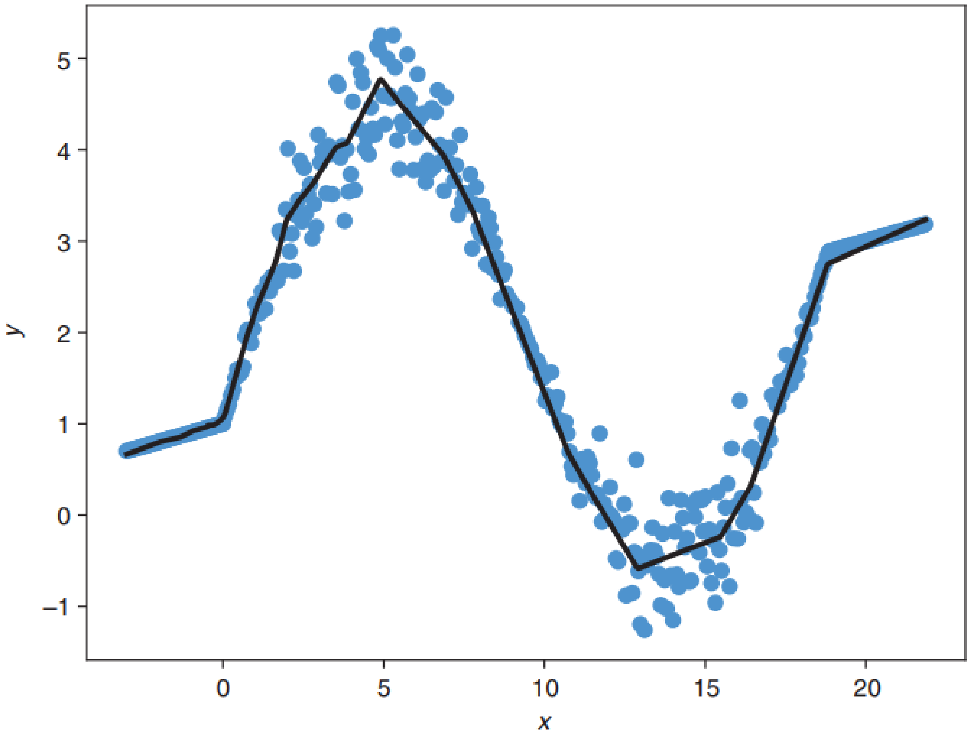
In the complicated case we still have x as input but
The output is going to be

$$(\mu_{x_i} = \text{f1 }(x;\boldsymbol{\beta}), \sigma^2_{x_i} = \text{f2 }(x;\boldsymbol{\beta}))$$



$$F(x) = (\mu_{x_i}, \sigma^2_{x_i})$$

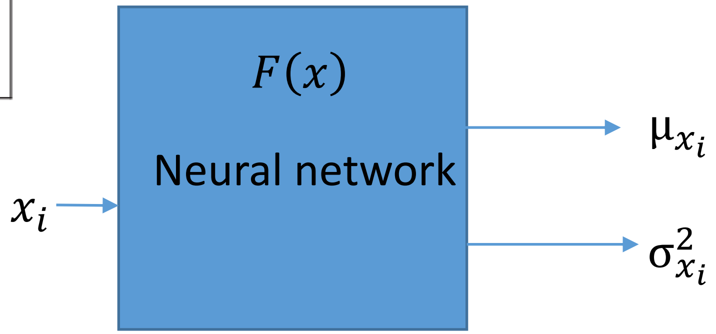
Introduction and motivation



$$Y_{x_i} \sim \mathrm{N}(\mu_{x_i} = \text{f1 } (\text{x};\boldsymbol{\beta}), \sigma^2_{x_i} = \text{f2 } (\text{x};\boldsymbol{\beta}))$$

In the complicated case we still have x as input but
The output is going to be

$$(\mu_{x_i} = \text{f1 } (\text{x};\boldsymbol{\beta}), \sigma^2_{x_i} = \text{f2 } (\text{x};\boldsymbol{\beta}))$$



$$F(x) = (\mu_{x_i}, \sigma^2_{x_i})$$

This network inputs x_i and outputs the parameter of a distribution
μ_{x_i} and $\sigma^2_{x_i}$

Formal introduction to DL

A *probabilistic model* is a model of the form

$$y = g(x; \boldsymbol{\beta}) + \varepsilon$$

ε~D where D is some distribution (say normal distribution)
β is a set of parameters that determine the model function

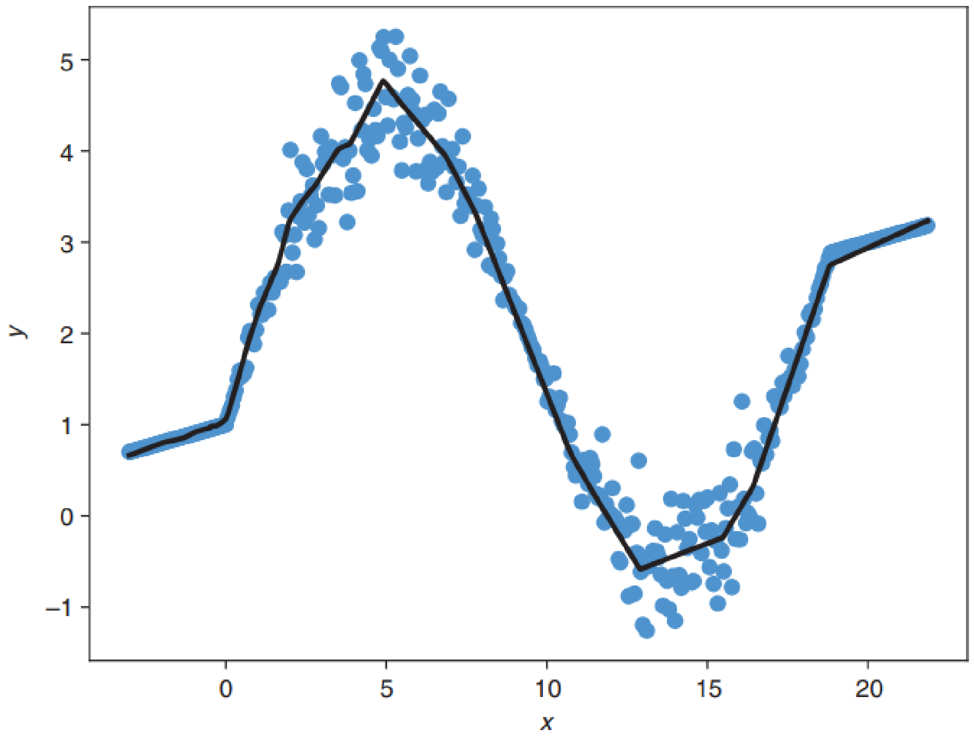
In other words, in a probability model : for every x we associate a distribution p(.|x) that depends on x and this distribution p(.|x) models the non-deterministic dependency of the random variable y on x.

Actually it is not hard to show that when $y = g(x; \boldsymbol{\beta}) + \varepsilon$ *then we have*

$$Y|X \sim D \Longleftrightarrow \varepsilon \sim D$$

Our goal is to learn a probability distribution $p_{\theta_x}(y|x)$ that best approximates $D = p_{data}(Y|X)$. $Here\ \theta_x$ is the parameter of the distribution $p_{\theta_x}(y|x)$.

Questions



Many questions :

(1) How do we choose F?
(2) How do we determine θ to fit the data?
(3) What if Y_{x_i} distribution is more complex?

$$Y_{x_i} \sim \mathrm{N}(\mu_{x_i} = \text{f1 (xi; θ)}, \sigma^2_{x_i} = \text{f2 (xi; θ)})$$

$$F(x; \theta) = (f_1(x; \theta), f_2(x; \theta))$$

$$f_1(x; \theta) = \mu_{x_i} \qquad f_2(x; \theta) = \sigma^2_{x_i}$$

Lets try to fit β 0 , β 1 with MLE

$one\ way\ to\ fit\ the\ model$

$Y_{x_i} \sim \mathrm{N}\left(\mu_{x_i} = \beta 0 + \beta 1 x_i \quad , \sigma_{x_i}^2 = \sigma^2\right)$ is via MLE :

$\mathrm{N}\left(\mu_{x_i} = \beta 0 + \beta 1 x_i \quad , \sigma_{x_i}^2 = \sigma^2\right)$ is a conditional normal distribution on x so we may write its pdf as follows :

g(yi; xi, β 0 , β 1)=$1/(\sqrt{2\pi\sigma} \quad e^{-\left(y_i - \mu_{x_i}\right)^2/\sigma^2}$

$$L(\beta 0 , \beta 1) = \prod_{i=1}^{N} \text{g(yi; xi, } \beta 0 , \beta 1\text{)} = \prod_{i=1}^{N} (1/(\sqrt{2\pi\sigma}\, e^{-\left(y_i - \mu_{x_i}\right)^2/\sigma_i^2}))$$

$it\ is\ usually\ easier\ to\ consider\ negative\ log\ \ lilkelhood\ function\ and\ to\ minimize\ neg$ log $lilkelhood$ instead Of maximizing L.

Now we model this more precisely

$one\ way\ to\ fit\ the\ model$

$$Y_{x_i} \sim \mathrm{N}\left(\mu_{x_i} = \beta\,0 + \beta\,1 x_i \quad , \sigma^2_{x_i} = \sigma^2\right) \text{ is via MLE :}$$

$\mathrm{N}\left(\mu_{x_i} = \beta\,0 + \beta\,1 x_i \quad , \sigma^2_{x_i} = \sigma^2\right)$ is a conditional normal distribution on x so we may write its pdf as follows :

g(yi; xi, $\beta\,0$, $\beta\,1$)=$1/(\sqrt{2\pi\sigma_i}\ e^{-\left(y_i - \mu_{x_i}\right)^2/\sigma^2}$

$$L(\beta\,0\,,\beta\,1) = \prod_{i=1}^{N} \text{g(yi; xi, } \beta\,0\,,\beta\,1\text{)=} = \prod_{i=1}^{N} (1/(\sqrt{2\pi\sigma}\ e^{-\left(y_i - \mu_{x_i}\right)^2/\sigma_i^2}\ ))$$

$it\ is\ usually\ easier\ to\ consider\ negative\ log\ \ lilkelhood\ function\ and\ to\ minimize\ neg$ log $lilkelhood$ instead Of maximizing L.

Now we model this more precisely

$so\ finding$ β 0 , β 1 that maximizes the function

$$L(\beta 0 , \beta 1) = \prod_{i=1}^{N} g(yi;\, xi,\, \beta 0 , \beta 1) = = \prod_{i=1}^{N} (1/(\sqrt{2\pi\sigma}\, e^{-\left(y_i - \mu_{x_i}\right)^2/\sigma^2}\ ))$$

$is\ the\ same\ problem\ as\ finding$ β 0 , β 1 that minimizes the function

$$-\log(L(\beta 0 , \beta 1)) = \sum_{i=1}^{n} -\log(\left(\frac{1}{\sqrt{2\pi\sigma}}\right)) + \left(y_i - \mu_{x_i}\right)^2/\sigma^2\ ))$$

$Hence:$

$$\left(\widehat{\beta 0}, \widehat{\beta 1}\right) = argmin_{(\beta 0\ ,\ \beta 1)}\ \sum_{i=1}^{n}(y_i - (\beta 0 + \beta 1 x_i))^2))$$

Now we model this more precisely

$Lets\ do\ the\ same\ thing\ but\ now\ we\ consider\ the\ more\ general\ model:$

$$\mathrm{N}\left(\mu_{x_i} = \text{f1 (xi; θ)}, \sigma^2_{x_i} = \text{f2 (xi; θ)}\right)$$

$lets\ find\ the\ parameter$ θ via MLE.

$\mathrm{N}\left(\mu_{x_i} = \text{f1 (xi; θ)}, \sigma^2_{x_i} = \text{f2 (xi; θ)}\right)$ is a conditional normal distribution on x so we may write its pdf as follows :

g(yi; xi,θ)=$1/(\sqrt{2\pi\sigma_i}\ e^{-\left(y_i-\mu_{x_i}\right)^2/\sigma_i^2}$

Note that $\sigma^2_{x_i}$ is now not a constant anymore

$$L(\theta) = \prod_{i=1}^{N} \text{g(yi; xi,θ)} = \prod_{i=1}^{N} (1/(\sqrt{2\pi\sigma}\ e^{-\left(y_i-\mu_{x_i}\right)^2/\sigma_i^2}))$$

$it\ is\ usually\ easier\ to\ consider\ negative\ log\ \ lilkelhood\ function\ and\ to\ minimize\ neg\log\ \ lilkelhood$ instead Of maximizing L.

Now we model this more precisely

$so\ finding$ θ that maximizes the function

$$L(\theta) = \prod_{i=1}^{N} \text{g(yi; xi,}\theta\text{)=} = \prod_{i=1}^{N} (1/(\sqrt{2\pi\sigma}\, e^{-\left(y_i - \mu_{x_i}\right)^2/\sigma_i^2}\ ))$$

$is\ the\ same\ problem\ as\ finding$ θ that minimizes the function

$$-\log(L(\theta)) = \sum_{i=1}^{n} -\log(\left(\frac{1}{\sqrt{2\pi\sigma}}\right)) + \left(y_i - \mu_{x_i}\right)^2/\sigma_i^2))$$

$Hence:$

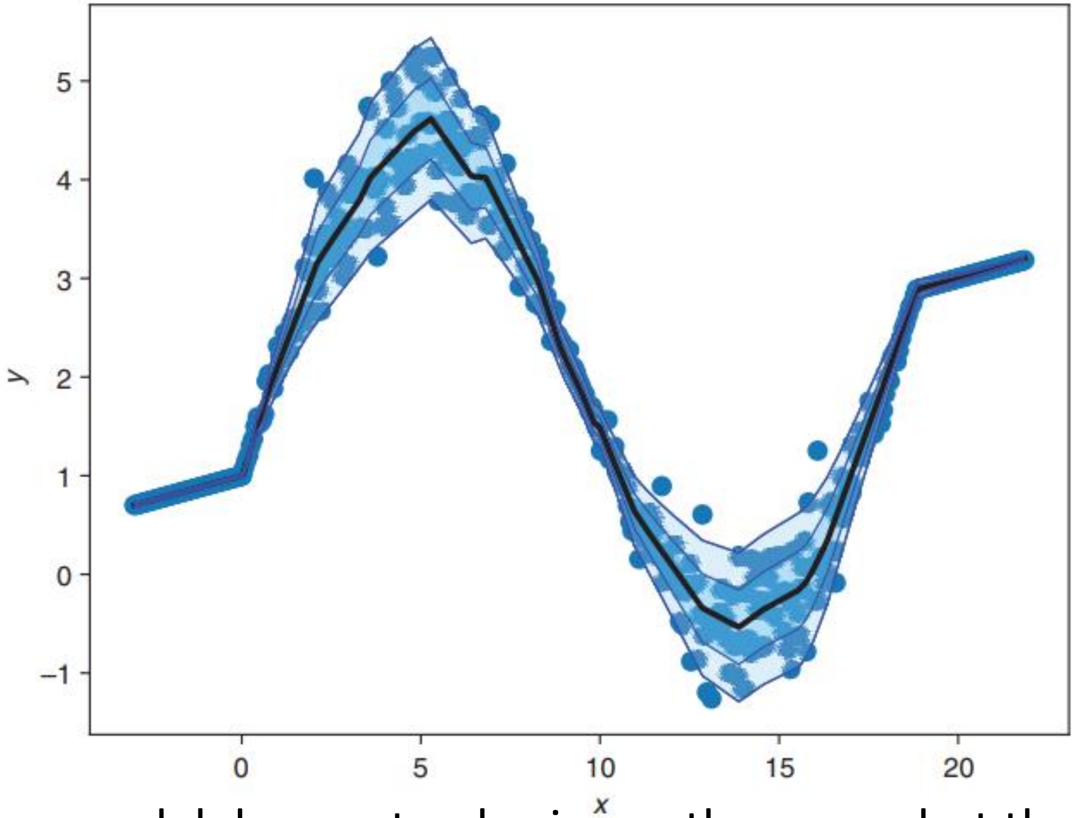
$$\hat{\theta} = argmin_{\theta} \sum_{i=1}^{n} -\log(\left(\frac{1}{\sqrt{2\pi\sigma}}\right)) + (y_i - f_1(x;\theta))^2/f_2(x;\theta)))$$

$how\ do\ we\ find$ $\hat{\theta}$? in general we do not need to worry about it. We optimize for $\hat{\theta}$ using some optimization software. All we have to do is provide the function that we want to optimize.

Neural Networks

$$\hat{\theta} = argmin_{\theta} \sum_{i=1}^{n} -\log(\left(\frac{1}{\sqrt{2\pi\sigma}}\right)) + (y_i - f_1(x;\theta))^2 / f_2(x;\theta)))$$

Minimizing the above function we can use it to fit the data :



With such a model, our model does not only give us the mean, but the standard deviation above every point!

Supervised Machine Learning

Lets try to formalize this.
NN(; **β**) (where **β is the parameter vector of the NN**)

such that NN(x, **β**)=θ_x where θ_x is the parameter that determine the distribution $p_{\theta_x}(y|x)$

Since the parameter **β** ultimately determines the parameters θ_x then the problem given in equation
Finding **β such that :**

$$p_{data}(Y|X) \approx p_{\boldsymbol{\beta}}(Y|X)$$

Given the above setup, we can find **β** by using MLE :

$$\boldsymbol{\beta} = argmmin_{\boldsymbol{\beta}}\, E_{(xi,yi)\sim p_{data}} -\log \ p_{\boldsymbol{\beta}}(yi|xi)$$

The above equation provides a **vast general principle**: most supervised ML falls in the above equation. In particular, most modern DL paradigm utilizes the above optimization scheme. Namely when p is normal we obtain regression problems, when p is categorical we obtain classification problems, etc.

Modeling general distribution

What if the relationship between x and y is not functional ? Example : on the top of every x, you have a multimodal distribution as in the following data :

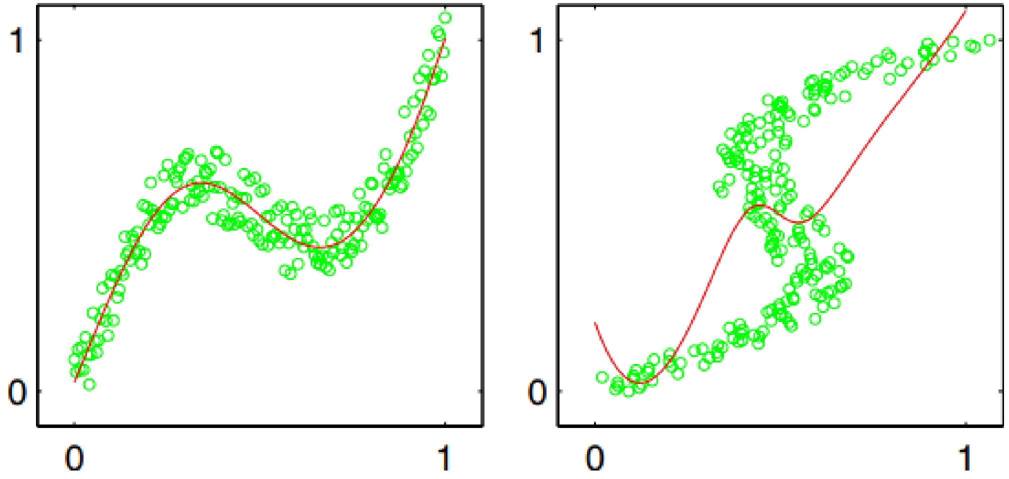
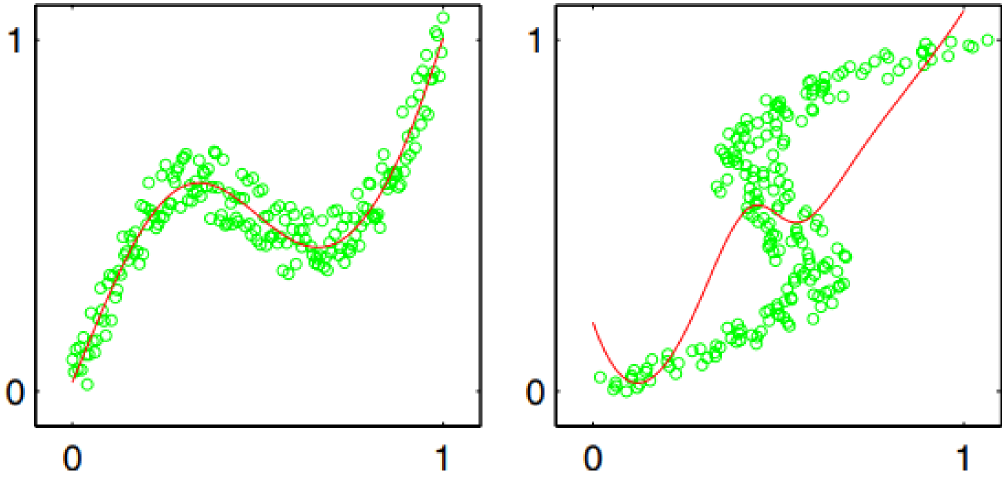


Image source :
http://users.isr.ist.utl.pt/~wurmd/Livros/school/Bishop%20-%20Pattern%20Recognition%20And%20Machine%20Learning%20-%20Springer%20%202006.pdf

Modeling general distribution

What if the relationship between x and y is not functional ? Example : on the top of every x, you have a multimodal distribution as in the following data :



In such cases we can use mixture models:
Mixture models are sum of Gaussians and they can be used to approximate any distribution

$$p(\mathbf{t}|\mathbf{x}) = \sum_{k=1}^{K} \pi_k(\mathbf{x}) \mathcal{N}\left(\mathbf{t}|\boldsymbol{\mu}_k(\mathbf{x}), \sigma_k^2(\mathbf{x})\right).$$

Image source :
http://users.isr.ist.utl.pt/~wurmd/Livros/school/Bishop%20-%20Pattern%20Recognition%20And%20Machine%20Learning%20-%20Springer%20%202006.pdf

Modeling general distribution

What if the relationship between x and y is not functional ? Example : on the top of every x, you have a multimodal distribution as in the following data :

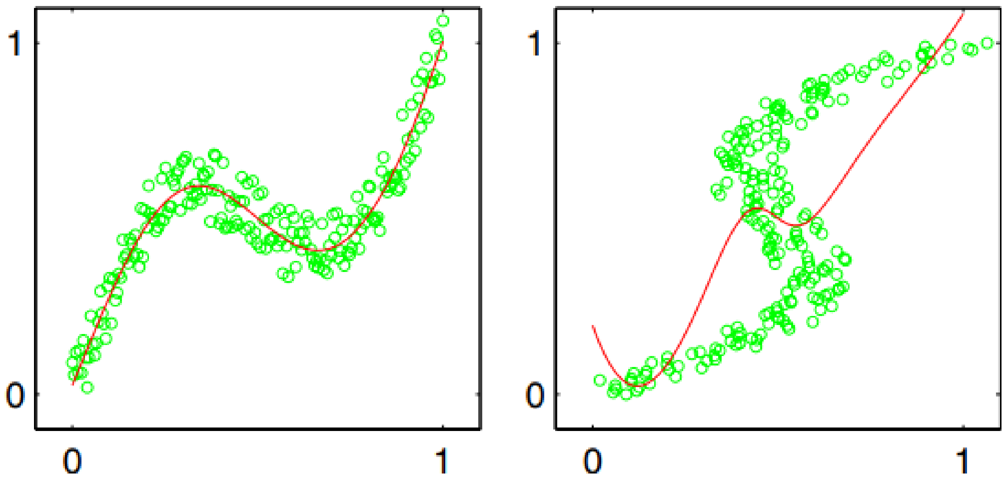
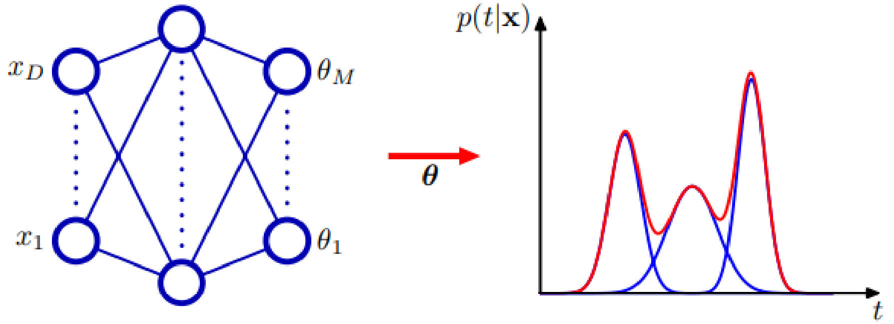


Image source : http://users.isr.ist.utl.pt/~wurmd/Livros/school/Bishop%20-%20Pattern%20Recognition%20And%20Machine%20Learning%20-%20Springer%20%202006.pdf

In such cases we can use mixture models:
Mixture models are sum of Gaussians and they can be used to approximate any distribution

$$p(\mathbf{t}|\mathbf{x}) = \sum_{k=1}^{K} \pi_k(\mathbf{x}) \mathcal{N}\left(\mathbf{t}|\boldsymbol{\mu}_k(\mathbf{x}), \sigma_k^2(\mathbf{x})\right).$$



In this case the NN outputs the 3K parameters of the distributions

$$\pi_k(\mathbf{x}_n, \mathbf{w}) \qquad \boldsymbol{\mu}_k(\mathbf{x}_n, \mathbf{w}) \qquad \sigma_k^2(\mathbf{x}_n, \mathbf{w})$$

Where K is the number of kernels in the Gaussian mixture model

Modeling general distribution

What if the relationship between x and y is not functional ? Example : on the top of every x, you have a multimodal distribution as in the following data :

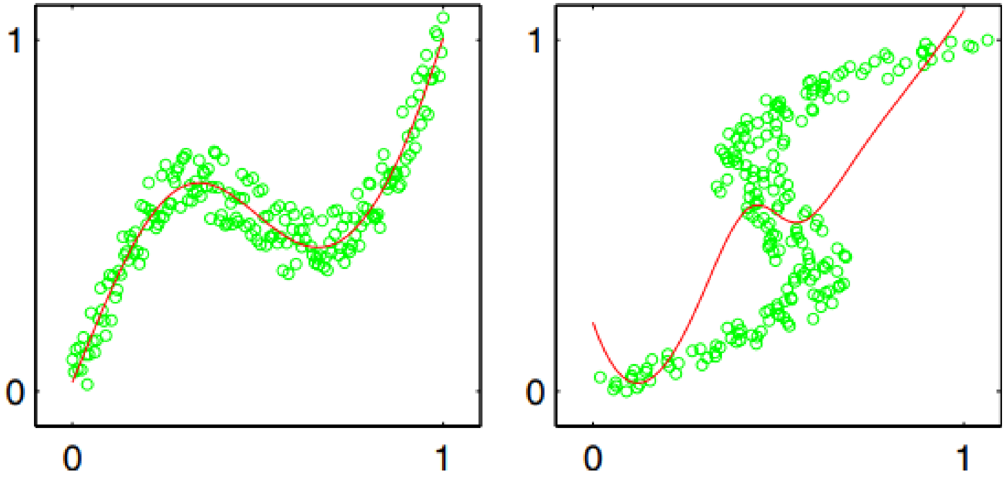
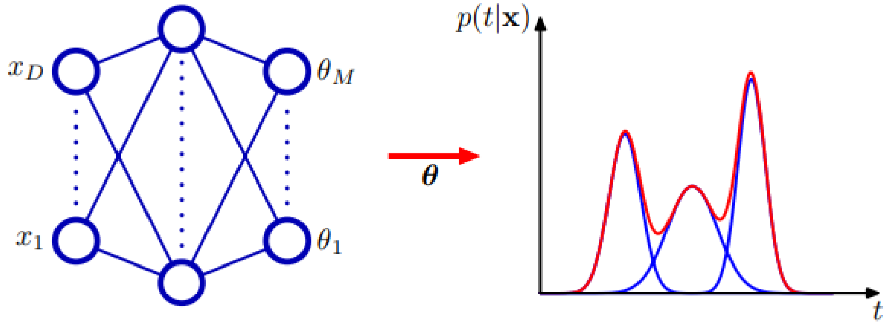


Image source : http://users.isr.ist.utl.pt/~wurmd/Livros/school/Bishop%20-%20Pattern%20Recognition%20And%20Machine%20Learning%20-%20Springer%20%202006.pdf

In such cases we can use mixture models:
Mixture models are sum of Gaussians and they can be used to approximate any distribution

$$p(\mathbf{t}|\mathbf{x}) = \sum_{k=1}^{K} \pi_k(\mathbf{x}) \mathcal{N}\left(\mathbf{t}|\boldsymbol{\mu}_k(\mathbf{x}), \sigma_k^2(\mathbf{x})\right).$$



It is not hard to see that negative logarithm of the likelihood is given by :

$$-\sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{k} \pi_k(\mathbf{x}_n, \mathbf{w}) \mathcal{N}\left(\mathbf{t}_n|\boldsymbol{\mu}_k(\mathbf{x}_n, \mathbf{w}), \sigma_k^2(\mathbf{x}_n, \mathbf{w})\right) \right\}$$