Lecture 2: Data, Variables and Distributions - 20/01/20

Binary Variables: Maximum Likelihood

Example: Coin Tossing

Suppose we have a biased coin. At each toss we get:

- Heads with probability μ
- Tails with probabilitu $1-\mu$

We toss the coin 100 times and observe 53 head and 47 tails. ***What is*** μ***?***

The Frequentist Answer: Pick the value of μ that makes the observations most probable.

Mathematically Encoding

- Each toss is a ***binary random variable*** X
- X can take two values: 1(head)/0(tail)
- For known, μ , X follows the ***Bernoulli Distribution*** :

$$p(x|\mu) = p(X=x|\mu) = \mu^x(1-\mu)^{(1-x)}$$

> $p(x|\mu)$ is the plausibility of a probability μ

- Date $D = \{x_1 = 1, x_2 = 1, x_3 = 0, \ldots\}$
- ***Assume we knew*** μ then:

$$p(D|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{(1-x_n)}$$

Frequentist View: find the maximum of this

Find Maximum

- Differentiate $p(D|\mu)$ is difficult
- But it is easier to find the maximum of $\ln p(D|\mu)$ and if $\mu^* = \arg\max_\mu f(\mu)$, then $\mu^* = \arg\max_\mu \ln f(\mu)$

> Using $f(x) = \prod_{n=1}^{N}[g(x)], \ln f(x) = \sum_{n=1}^{N}[\ln g(x)]$

$$\ln p(D|\mu) = \sum_n [x_n \ln\mu + (1-x_n)\ln(1-\mu)]$$

> Using $f(\mu) = x\ln\mu, \frac{d}{d\mu}f = \frac{x}{\mu}$

$$\frac{d}{d\mu}\ln p(D|\mu) = \sum_{n=1}^{N}[\frac{x_n}{\mu} - \frac{1-x_n}{1-\mu}]$$

Let $\frac{d}{d\mu}\ln p(D|\mu) = 0$

$$\sum_N[\frac{x_n}{\mu} - \frac{1-x_n}{1-\mu}] = 0$$
$$\sum_N \frac{x_n}{\mu} = \sum_N \frac{1-x_n}{1-\mu}$$
$$\frac{1}{\mu}\sum_N x_n = \frac{1}{1-\mu}\sum_N[1-x_n]$$
$$\frac{1}{\mu}\sum_N x_n = \frac{N}{1-\mu} - \frac{1}{1-\mu}\sum_N x_n$$
$$\frac{1}{\mu\,\cancel{(1-\mu)}}\sum_N x_n = \frac{N}{\cancel{1-\mu}}$$

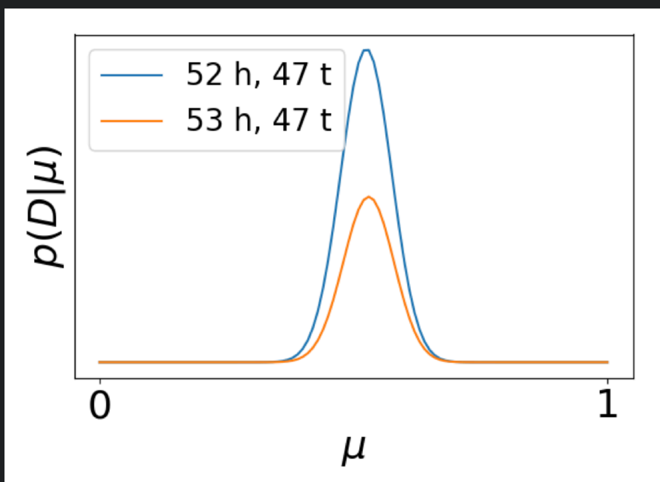
> *Maximiser:*
>
> $$\mu_{ML} = \frac{\sum_N x_n}{N}$$

Terminology

- $p(D|\mu)$ is the ***joint probability*** of D
- $p(D|\mu)$ is also the ***likelihood*** of μ
- In $p(D|\mu)$ is the ***log-likelihood*** of μ
- μ_{ML} Is the ***maximum-likelihood parameter***
- As all $x_n \in D$ are drawn independently from the same distribution we say they are ***independent and identically distributed*** or ***i.i.d***

Problems with Maximum-Likelihood

- If we got two heads, $\mu = 1$ is yielded, but not sensible
- μ might not be a single answer
- Insufficient data leads to uncertainty about μ
- Taking uncertainty into account leads to better reasoning



> $p(D|\mu)$ is not a probability distribution over μ as the area under the curve isn't 1

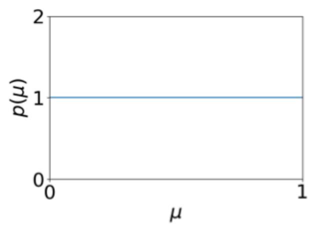
Bayesian Variables: A Bayesian Approach

[如何通俗理解Beta分布 - 知乎]

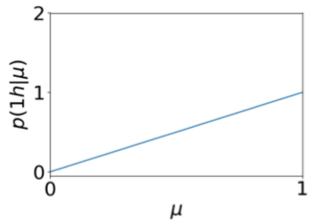
> $posterior \propto likelihood \times prior$

Example: first coin is head and second coin is tail

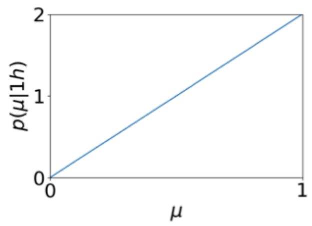
Start with a prior distribution for μ.
Here we use the uniform distribution.



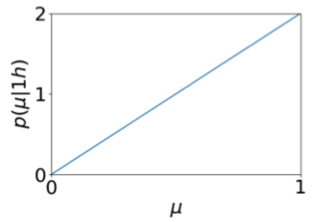
Likelihood $p(1h|\mu)$ charts probability of a head given μ.



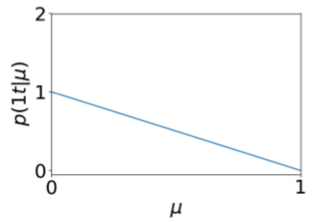
For each μ (pointwise) multiply $p(\mu)$ with $p(1h|\mu)$, then rescale (normalise) so the integral sums to 1.
This is the posterior probability density.



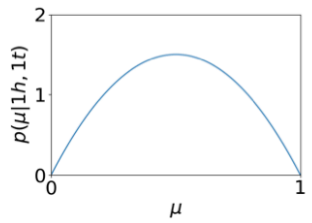
The new prior is $p(\mu|1h)$ (the posterior from the previous toss).

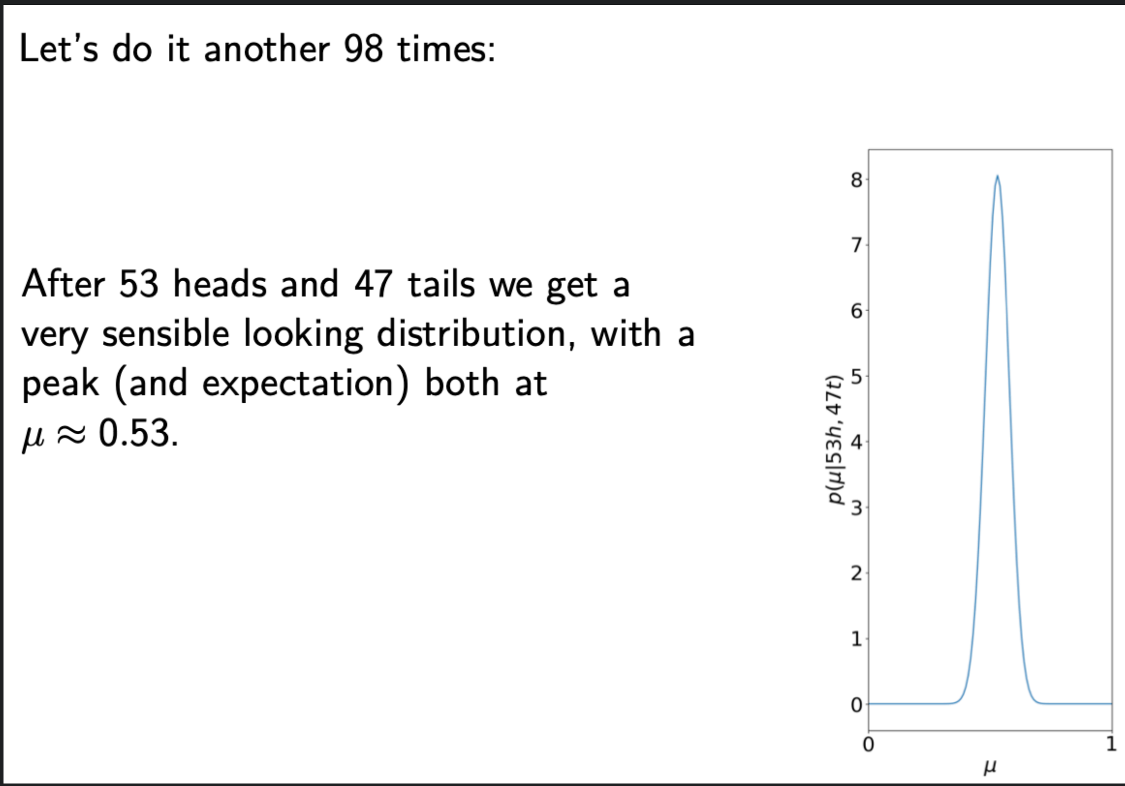


Likelihood $p(1t|\mu)$ charts probability of a tail given μ.



Pointwise multiply prior and likelihood then renormalise to get the posterior.





The Beta Distribution

> The beta distribution describes continuous random-variables in the range $[0, 1]$ and has the form:
>
> $$Beta(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{(a-1)}(1-\mu)^{(b-1)}$$
>
> > ***Note that***: $\Gamma(n) = (n-1)! \forall n \in \mathbb{N}^{+}$
>
> Ignore the normalisation:
>
> $$Beta(\mu|a,b) \propto \mu^{(a-1)}(1-\mu)^{(b-1)}$$
>
> Mean:
>
> $$E(\mu|a,b) = \frac{a}{a+b}$$
>
> Variance:
>
> $$var(\mu|a,b) = \frac{ab}{(a+b)^2(a+b+1)}$$

A Beta Prior

> We choose a ***Beta prior*** $p(\mu) = Beta(\mu|a,b) = \mu^{(a-1)}(1-\mu)^{(b-1)}$ for our coin then observe m heads and l tails

> *For uncluttered notation, we sometimes wirte $p(\mu) = p(\mu|a,b)$*

Using Bayes Theory our ***posterior*** looks like:

$$\begin{aligned} p(\mu|D) &\propto p(D|\mu)p(\mu) \\ &\propto \mu^m(1-\mu)^l\mu^{(a-1)}(1-\mu)^{(b-1)} \\ &= \mu^{(m+a-1)}(1-\mu)^{(l+b-1)} \end{aligned}$$

Once correctly ***normalised***:

$$p(\mu|D) = Beta(\mu|m+a, l+b)$$

> New observation added to the experience data

> When our posterior takes the same form as the prior, then the prior is said to be ***conjugate***

Conjugate Beta Priors

With Beta prior $Beta(\mu|a,b)$ we observe m heads and l tails.

> ***Prior Estimate:*** $E[\mu|a,b] = \frac{a}{a+b}$
>
> ***Posterior Estimate:*** $E[\mu|a,b,m,l] = \frac{a+m}{a+m+b+l}$

Recall that $Beta(\mu|a,b) \propto \mu^{(a-1)}(1-\mu)^{(b-1)}$

- Can interpret a and b as ***effective prior observations***
- $a=1$ and $b=1$ give a ***flat prior*** ($p(\mu)$ constant)
- a and b must be greater than 0
- a and b don't necessarily need to be integers

Reak Valued Data

Another form of data we deal with regularly is unbounded reals ($x_n \in \mathbb{R}$). Can often model this with a ***Gaussian***:

- the Gaussian has many nice properties (as well will see)
- reasons to expect data to be (approximately) Gaussian

> a Gaussian prior can induce a Caussian posterior (conjugacy)

- we can test data for its ***Gaussianity***

The Gaussian (Normal) Distribution

$$p(x|\mu,\sigma^2) = N(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp[-\frac{1}{2\sigma^2}(x-\mu)^2]$$

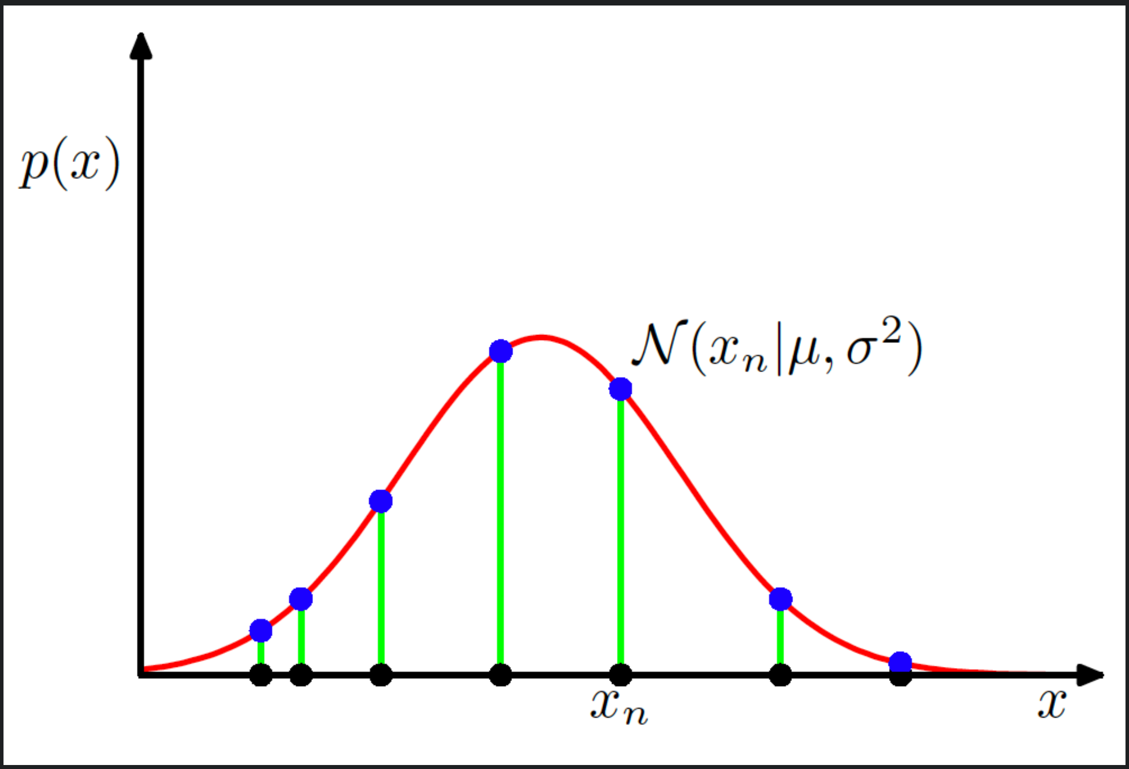
for real values random samples $x_n \in \mathbb{R}$

Properties:

- $N(x|\mu,\sigma^2) > 0$ for $x \in \mathbb{R}$
- $\int_{-\infty}^{\infty} N(x|\mu,\sigma^2)dx = 1$
- $E[x] = \mu$
- $var[x] = \sigma^2$

Gaussian Likelihood

If we draw N samples $\mathbf{x} = (x_1, \dots, x_N)^T$ ***i.i.d*** from our Gaussian, e.g. $x_n \sim N(x|\mu, \sigma^2)$



The ***likelihood*** is:

$$
\begin{aligned}
p(\mathbf{x}|\mu, \sigma^2) &= \prod_{n=1}^{N} N(x_n|\mu, \sigma^2) \\
&= \prod_{n=1}^{N} [\frac{1}{\sqrt{2\pi\sigma^2}} \exp[-\frac{1}{2\sigma^2}(x_n - \mu)^2] \\
&= (\frac{1}{\sqrt{2\pi\sigma^2}})^N \exp[-\sum_{n=1}^{N} \frac{1}{2\sigma^2}(x_n - \mu)^2] \\
&= (\frac{1}{\sqrt{2\pi\sigma^2}})^N \exp[-\frac{1}{2\sigma^2}[\sum_n x_n^2 - \sum_n 2\mu x_n + N\mu^2]]
\end{aligned}
$$

Maximum likelihood parameters are a pair of values $\mu = \mu_{ML}, \sigma^2 = \sigma^2_{ML}$ that maximises the likelihood.

Easier to find maximisers using ***log likelihood***:

$$
\begin{aligned}
\ln p(\mathbf{x}|\mu, \sigma^2) &= N \ln(\frac{1}{\sqrt{2\pi\sigma^2}}) + \ln(\exp[-\frac{1}{2\sigma^2}[\sum_n x_n^2 - \sum_n 2\mu x_n + N\mu^2]) \\
&= -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}[\sum_n x_n^2 - \sum_n 2\mu x_n + N\mu^2] \\
&= -\frac{N}{2} \ln \sigma^2 - \frac{N}{2} \ln(2\pi) - \frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2
\end{aligned}
$$

Differentiate $\ln p(\mathbf{x}|\mu, \sigma^2)$ and set to zero:

$$
\begin{aligned}
\frac{d}{d\mu}\ln p(\mathbf{x}|\mu,\sigma^2) &= -\frac{1}{2\sigma^2}\sum_{n=1}^{N}[-2x_n+2\mu] \\
let\ \frac{d}{d\mu}\ln p(\mathbf{x}|\mu,\sigma^2) &= 0 \\
-\frac{1}{2\sigma^2}\sum_{n=1}^{N}[-2x_n+2\mu] &= 0 \\
\sum_{n=1}^{N}\mu &= \sum_{n=1}^{N}x_n \\
\mu &= \frac{1}{N}\sum_{n=1}^{N}x_n
\end{aligned}
$$

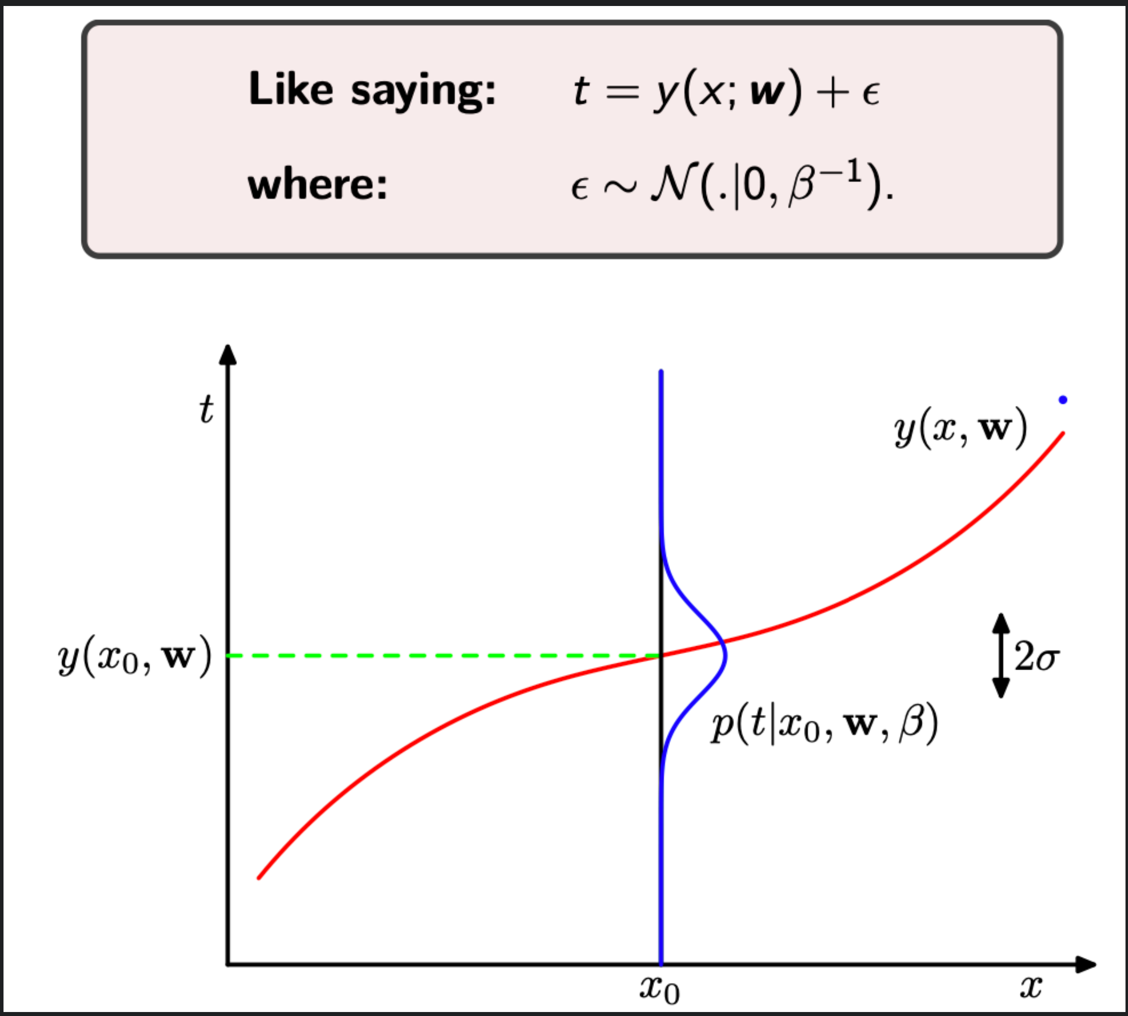
$$
\begin{aligned}
\frac{d}{d\sigma^2}\ln p(\mathbf{x}|\mu,\sigma^2) &= -\frac{N}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{n=1}^{N}(x-M-\mu)^2 \\
let\frac{d}{d\sigma^2}\ln p(\mathbf{x}|\mu,\sigma^2) &= 0 \\
\frac{N}{2\sigma^2} &= \frac{1}{2\sigma^4}\sum_{n=1}^{N}(x_n-\mu)^2 \\
\sigma^2 &= \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu)^2
\end{aligned}
$$

Thus

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

Maximum Likelihood Curve Fitting

> ***Why assume Gaussian?*** Gaussian Distribution is a convinient default assumption, it makes the maths easy



Using the curve fitting example from ***Lecture 1***:

- N inputs $\mathbf{x} = (x_1, \ldots, x_N)^T$
- N targets $\mathbf{t} = (t_1, \ldots, t_N)^T$

Assume given x_i, then t_i ***Gaussian*** with mean $y(x_i; \mathbf{w})$ (a polynomial with weights $\mathbf{w}$), i.e.

$$p(t_i|x_i, \mathbf{w}, \beta) = N(t_i|y(x_i; \mathbf{w}), \beta^{-1})$$

> ***Note that*** : $f(x; a, b, c)$ means x is the independent variable while a, b, c are the parameters

Where

$$y(x_i; \mathbf{w}) = \sum_{j=0}^{M} w_j x_i^j$$

In which M is the order of the hypothesis function.

β is the ***precision*** (inverse variance), $\beta^{-1} = \sigma^2$

Using $\{\mathbf{x}, \mathbf{t}\}$ to find the maximum likelihood parameters for $\mathbf{w}$ and β. The likelihood:

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}) = \prod_{n=1}^{N} N(t_n|y(x_n; \mathbf{w}), \beta^{-1})$$

> ***Note that***: for this Gaussian distribution, the mean is $y(x_n; \mathbf{w})$ and the variance β^{-1} , and random variable being t_n

$$\ln p(\mathbf{t}|\mathbf{x},\mathbf{w}) = -\frac{\beta}{2}\sum_{n=1}^{N}[y(x_n;\mathbf{w})-t_n]^2 + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi)$$

which, for $\mathbf{w}$, is the same as minimising:

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}[y(x_n,\mathbf{w})-t_n]^2$$

Insights

- maximising likelihood (minimising error) gives $\mathbf{w}_{ML}$ (independent of β)
- can then find β_{ML}
- now have predictive distribution for new samples t given x : $p(t|x,\mathbf{w}_{ML},\beta_{ML}) = N(t|y(x;\mathbf{w}_{ML}),\beta_{ML})$, this measures uncertainty over t
- doesn't measure uncertainty in $\mathbf{w}$ or β

Gaussian: A More Bayesian View

Moving Towards a More *Bayesian* View

Assume samples $\mathbf{x} = (x_1,\ldots,x_n)^T$ from $N(x_i|\mu,\sigma^2)$

- Assume we know the variance σ^2
- Define prior on $\mu, p(\mu)$
- But what shold the ***prior*** look like?

Likelihood:

$$\begin{aligned} p(\mathbf{x}|\mu) &= \prod_{n=1}^{N} N(x_n|\mu,\sigma^2) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}^N}\exp[-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2] \end{aligned}$$

> Since the prior and the posterior should be conjugate

Using ***Bayes Rule***, the posterior:

$$p(\mu|\mathbf{x}) \propto p(\mathbf{x}|\mu)p(\mu)$$

and the ***likelihood*** has form:

$$p(\mathbf{x}|\mu) \propto \exp[-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2]$$

> This is an exponential that is (-ve) quadratic in μ since the index is $-\frac{1}{2\sigma^2}[\sum_n x_n^2 - \sum_n 2\mu x_n + N\mu^2]$

So $p(\mu)$ should also have similar form:

$$\begin{aligned} p(\mu) &\propto \exp[-\frac{(\mu-m_0)^2}{2s_0^2}] \\ &\propto N(\mu|m_0,s_0^2) \end{aligned}$$

> ***Prior*** is a ***Gaussian***

From conjugacy, we know the ***posterior*** will also be Gaussian, so

$$p(\mu|\mathbf{x}) = N(\mu|m_N, s_N^2)$$

By working through the maths:

$$\begin{aligned} m_N &= \frac{\sigma^2}{Ns_0^2 + \sigma^2} m_0 + \frac{Ns_0^2}{Ns_0^2 + \sigma^2} \mu_{ML} \\ \frac{1}{s_N^2} &= \frac{1}{s_0^2} + \frac{N}{\sigma^2} \end{aligned}$$

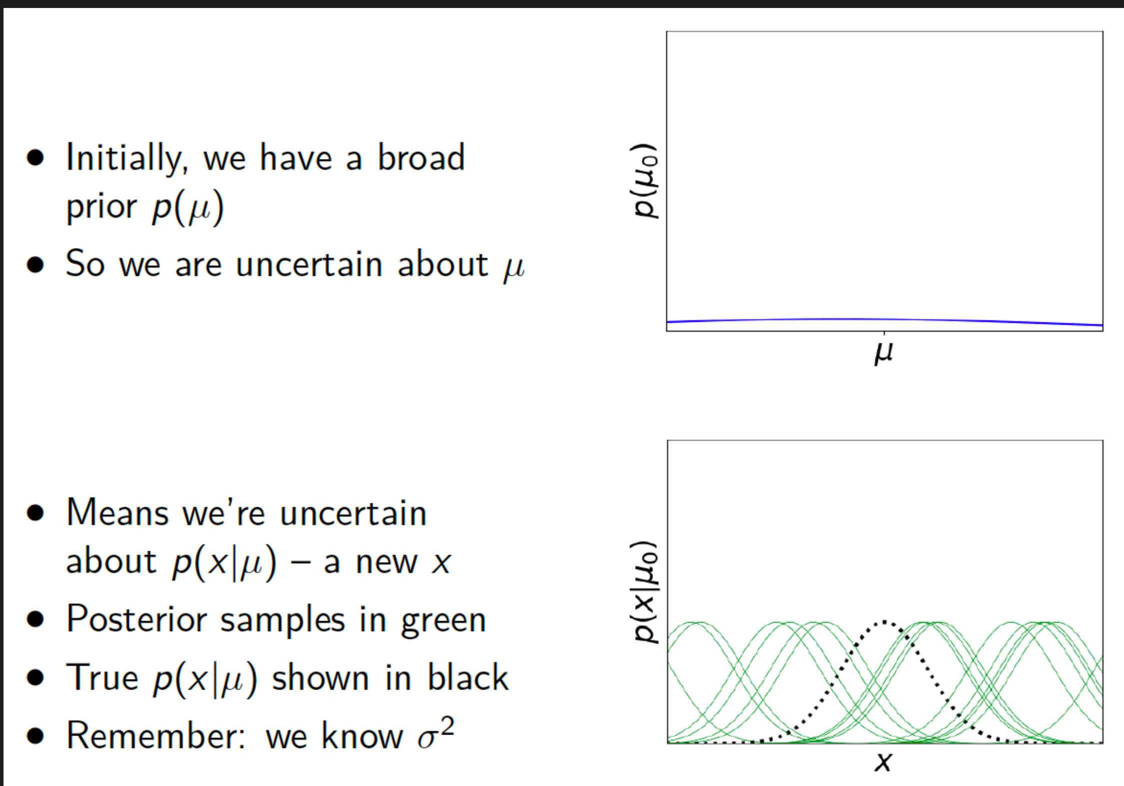
where μ_{ML} is the ***maximum likelihood mean*** from before:

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

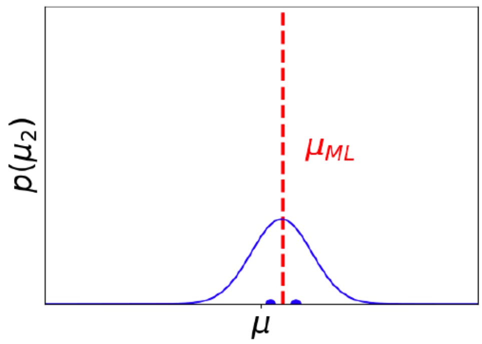
Insight

- m_N is compromise between m_0 and μ_{ML}
- For large N, $m_N \approx \mu_{ML}$ (refer to ***Lecture 1***)
- Precisions, e.g. $\frac{1}{\sigma^2}$, more natural than variance
- For large N, variance of estimate vanishes, i.e. $s_N^2 \approx 0$
- As $s_0^2 \to \infty$, then $s_N^2 \to \frac{\sigma^2}{N}$

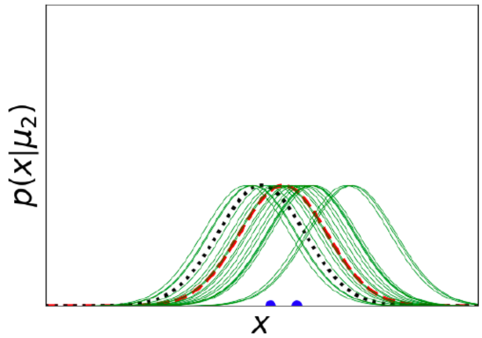
Machine Learning Application



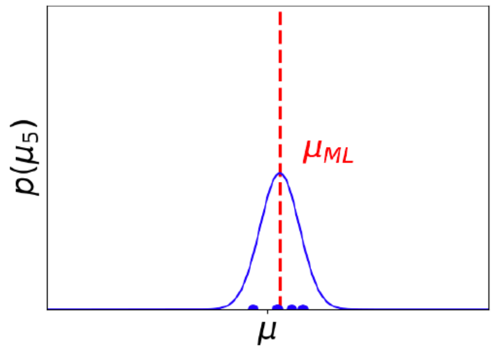
- After observing 2 samples, we know more about μ
- Captured by $p(\mu|\boldsymbol{x})$



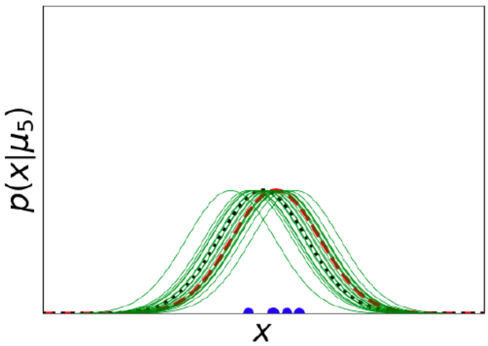
- Meaning we are a little more sure about $p(x|\mu)$
- $p(x|\mu_{\text{ML}})$ shown in red



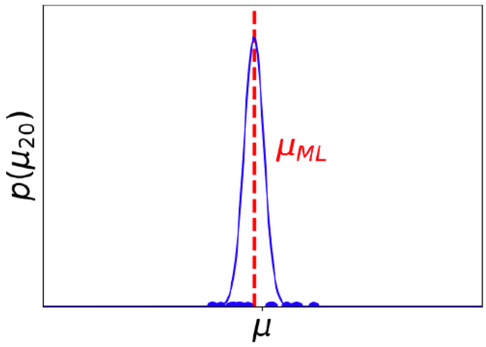
- Adding 3 more samples narrows our uncertainty about μ
- $p(\mu|\boldsymbol{x})$ more peaked around μ_{ML}



- This makes us more certain about $p(x|\mu)$
- But there is still uncertainty not captured by μ_{ML}



- Another 15 samples and $p(\mu|\boldsymbol{x})$ is sharply peaked around μ_{ML}



- More certain now about $p(x|\mu)$ (and accurate)
- μ_{ML} is closer to the true mean too