PROBABILISTIC MODELS

LECTURE 2
SECTION 1
JUNE 4TH

INSTITUTE FOR APPLIED
COMPUTATIONAL SCIENCE
AT HARVARD UNIVERSITY

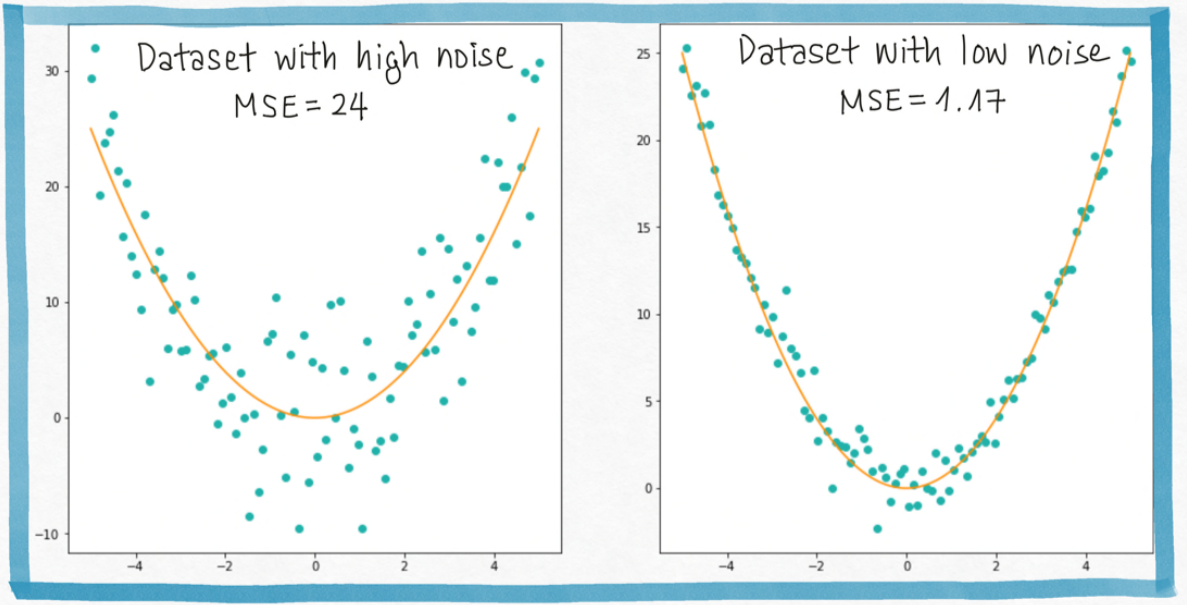
UNIVERSITY of
RWANDA

PROBALISTIC MODELS

REASONING ABOUT NOISE:

In Lecture 1 we saw that a model that captures the trend perfectly can still have high MSE due to the noise in the data.

But how can we explicitly account for noise in our model?



PROBABILISTIC MODELS:

A probabilistic model for regression is:

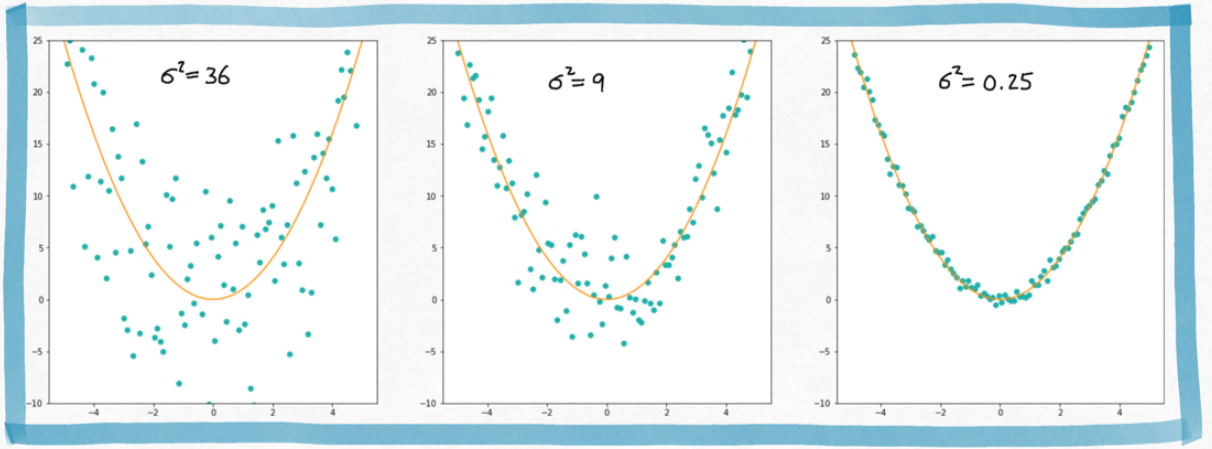
$$y = f_w(x) + \varepsilon$$

where ε is a random variable. This says that the observed target differs from the predicted $f(x)$ by a random amount of noise ε.

In regression, we assume that the noise ε is normally distributed

$$\varepsilon \sim N(0, \sigma^2)$$

this means that on average the noise is zero, the "noise level" in the data is given by σ^2.



THE LIKELIHOOD FUNCTION:

In the probabilistic model, $y = f_w(x) + \varepsilon$, y is a random variable. Because $\varepsilon \sim N(0, \sigma^2)$, y is also normally distributed:

$$y \sim N(f_w(x), \sigma^2)$$

We write the distribution of y as $p(y|x,w)$, this is called the likelihood.

$P(y|x, w)$ gives the likelihood of observing a y for a given x and a given model f_w.

Example: compute the likelihood of each observation $(1,5)$ & $(1,6)$ under the model $y = (2x+4) + \varepsilon$, $\varepsilon \sim N(0, 0.5)$.

The likelihd: $p(y|x,w) = N(y; 2x+4, 0.5) = \frac{1}{\sqrt{2\pi}\,0.5} \exp\left\{-\frac{(2x+4-y)^2}{2 \cdot 0.5}\right\}$

plug in (1,5): $p(y=5|x=1,w) = \frac{1}{\sqrt{2\pi}\,0.5} \exp\left\{-\frac{(2\cdot 1+4-5)^2}{2 \cdot 0.5}\right\} = 0.29$

plug in (1,6): $p(y=6|x=1,w) = \frac{1}{\sqrt{2\pi}\,0.5} \exp\left\{-\frac{(2\cdot 1+4-6)^2}{2 \cdot 0.5}\right\} = 0.8$

Under the model, $(1,5)$ is less likely than $(1,6)$ $(1,5)$ is a more exceptional point and $(1,6)$ is more typical.

INFERENCE

THE MAXIMUM LIKELIHOOD PRINCIPLE:

The likelihood function can be used to select a "best" model:

The Maximum Likelihood Principle
We should select a model that maximizes the likelihood of the observed data.

EXAMPLE: Suppose we have observed that a house with size $x = 70\,m^2$ selling for $y = \$500,000$. Suppose a model class:

$$y = w_1 x + w_0 + \varepsilon, \quad \varepsilon \sim N(0, 2000)$$

Let's compare two models:

M1: $y = 5{,}000x + 100{,}000$ **M2**: $y = 10{,}000x + 10$

For **M1**: $\log p(y|x, w_1 = 5{,}000, w_0 = 100{,}000) = \log \frac{1}{\sqrt{2\pi \cdot 2000}} + \log \cdot \exp\left\{ - \frac{(500{,}000 - (5000 \cdot 70 + 100{,}000))^2}{2 \cdot 2000} \right\} = -321$

For **M2**: $\log p(y|x, w_1 = 10{,}000, w_0 = 10) = \log \frac{1}{\sqrt{2\pi \cdot 2000}} + \log \exp\left\{ - \frac{(500{,}000 - (10{,}000 \cdot 70 + 10))^2}{2 \cdot 2000} \right\} = -5009$

This means that under M1 the data we observed is <u>more likely</u>, under M2 the data we we observed is <u>extremely unlikely</u>.

By the maximum likelihood principle, we prefer **M1** over **M2**.

MAXIMIZING THE LIKELIHOOD:

Optimal values of the likelihood occur at stationary points of the gradient:

$$\nabla_w \mathcal{L}(w) = \nabla_w \prod_{n=1}^{N} N(y_n; f_w(x_n), \sigma^2) = 0$$

Equivalently, we can find where the gradient of the log-likelihood is zero:

$$\nabla_w \ell(w) = \nabla_w \log \prod_{n=1}^{N} N(y_n; f_w(x_n), \sigma^2) = \nabla_w \sum_{n=1}^{N} \log N(y_n; f_w(x_n), \sigma^2) = 0$$

We get:

$$\begin{aligned}
\nabla_w \ell(w) &= \sum_{n=1}^{N} \nabla_w \log N(y_n; f_w(x_n), \sigma^2) \\
&= \sum_{n=1}^{N} \nabla_w \left[\log \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) + \log \exp \left\{ - \frac{(y_n - f_w(x_n))^2}{2\sigma^2} \right\} \right] \\
&= \sum_{n=1}^{N} \nabla_w \left[- \frac{(y_n - f_w(x_n))^2}{2\sigma^2} \right] \\
&= - \frac{N}{2\sigma^2} \nabla_w \underbrace{\frac{1}{N} \sum_{n=1}^{N} (y_n - f_w(x_n))^2}_{\text{MSE}(w)} = 0
\end{aligned}$$

We see that $\nabla_w \ell(w) = 0$ whenever $\nabla_w \text{MSE}(w) = 0$. So:

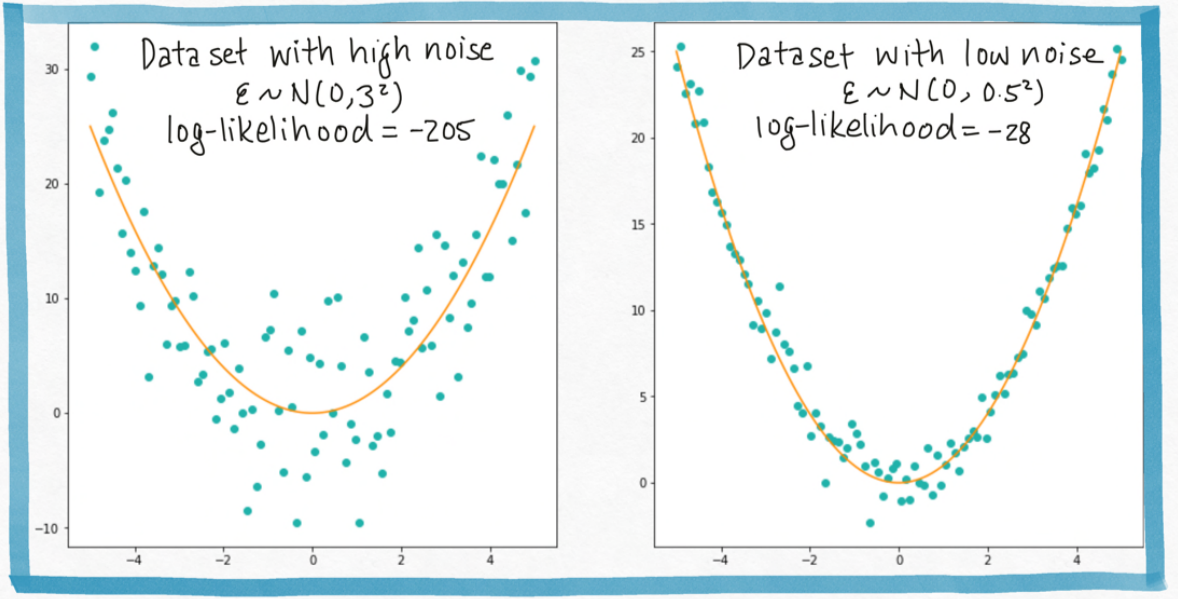
Maximizing log-likelihood is the same as minimizing MSE

EVALUATING PROBABILISTIC MODELS

UNDERFITTING OR HIGH NOISE?

An ideal model will have similar log-likelihood on train data as on test data. Both log-likelihoods should be high.

Is this always true?



Log likelihood is not a reliable way to detect underfitting!

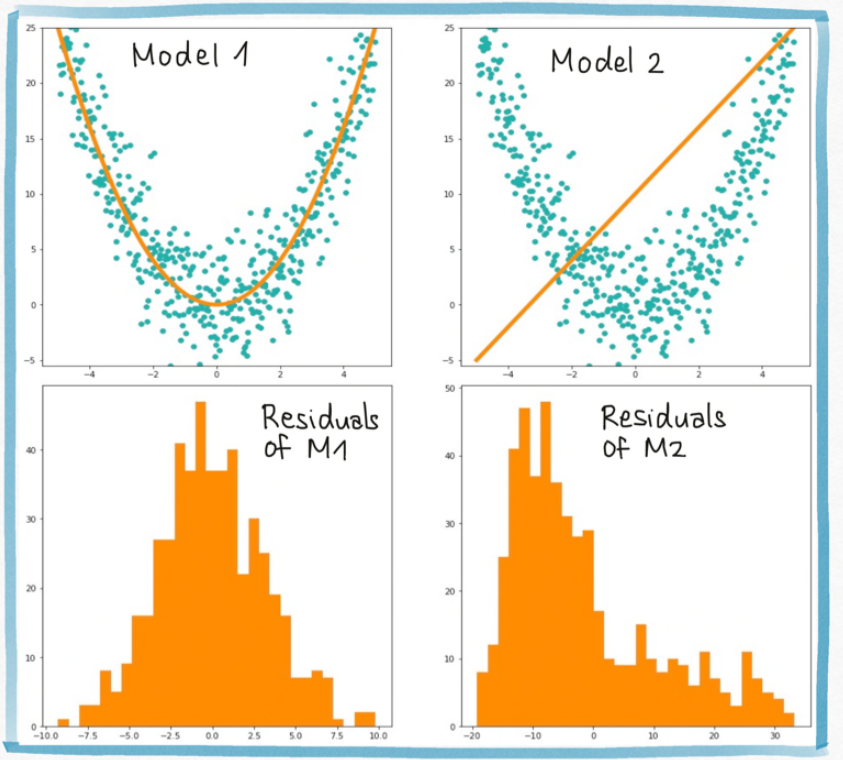
So what is the point of probabilistic models?

EVALUATION: RESIDUAL HISTOGRAM

If our probabilistic model $y_n = f_w(x_n) + \varepsilon_n$, $\varepsilon_n \sim N(0, \sigma^2)$ is correct, then our residuals $y_n - f_w(x_n) = \varepsilon_n$ will be normally distributed.

Histograms of the residuals show us incorrect modeling assumptions:

- maybe the model class f_w is incorrect
- maybe the noise model $\varepsilon_n \sim N(0, \sigma^2)$ is incorrect.

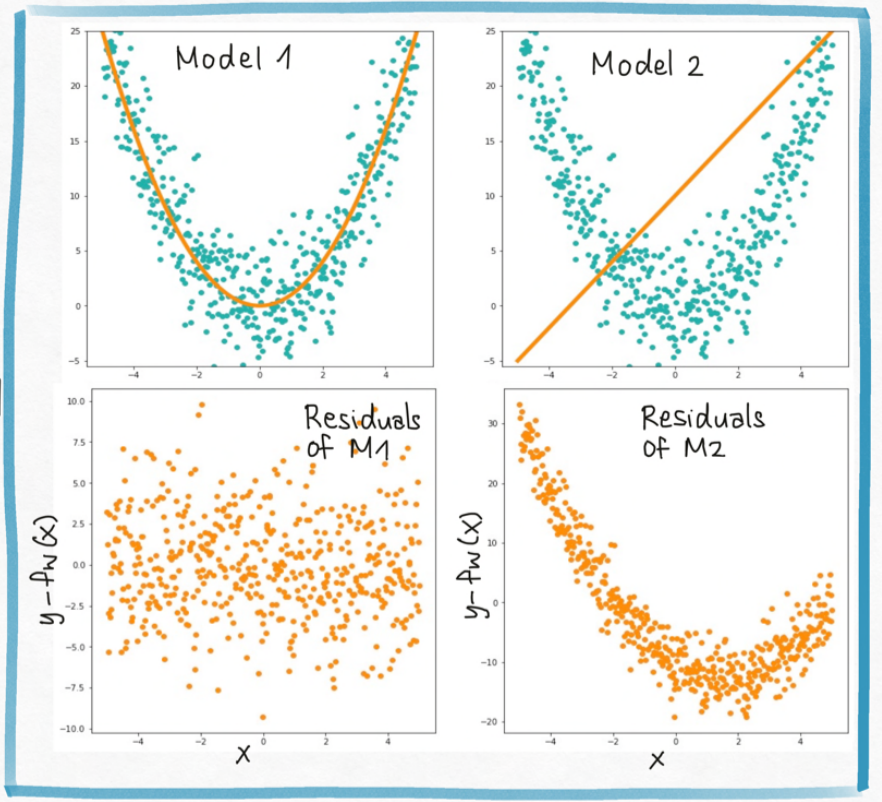


EVALUATION: RESIDUALS VS X

If our probabilistic model $y_n = f_w(x_n) + \varepsilon_n$, $\varepsilon_n \sim N(0, \sigma^2)$ is correct, then our residuals $y_n - f_w(x_n) = \varepsilon_n$ will be independent from x.

Scatter plots of the residuals show us incorrect modeling assumptions:

- maybe the model class f_w is incorrect
- maybe the noise model $\varepsilon_n \sim N(0, \sigma^2)$ is incorrect.



EVALUATION: RESIDUALS VS X

If our probabilistic model $y_n = f_w(x_n) + \varepsilon_n$, $\varepsilon_n \sim N(0, \sigma^2)$ is correct, then our residuals $y_n - f_w(x_n) = \varepsilon_n$ will be independent from x.

Scatter plots of the residuals show us incorrect modeling assumptions:

- maybe the model class f_w is incorrect
- maybe the noise model $\varepsilon_n \sim N(0, \sigma^2)$ is incorrect.