06-20416 and 06-12412
(Intro to) Neural Computation

03 – Maximum Likelihood

Per Kristian Lehre

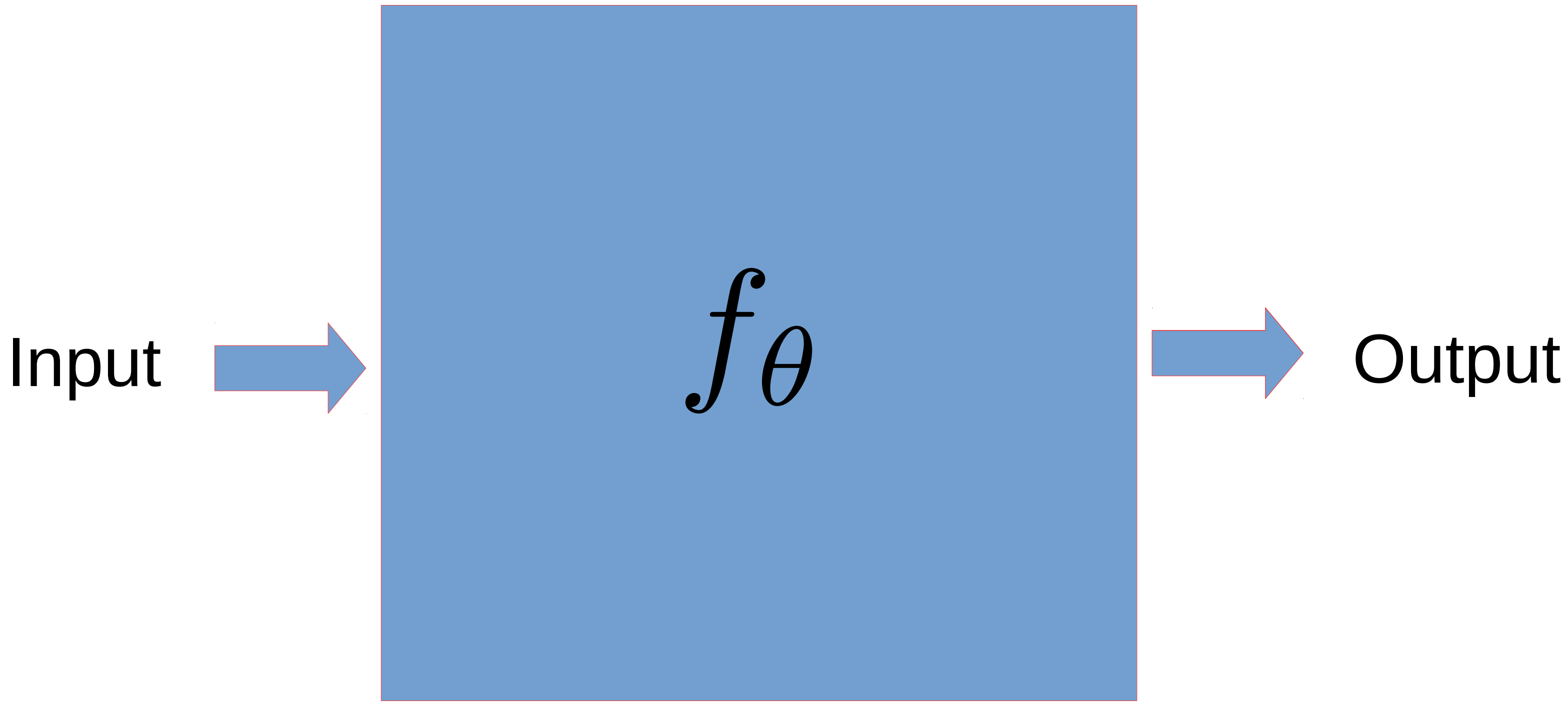
Previous lecture

- Linear regression models
  - model *linear* relationship between input and output
  - Mean square error as cost function
- Optimisation
- Derivatives
  - The chain rule
- Ordinary Least Square (OLS)
- Gradient Descent

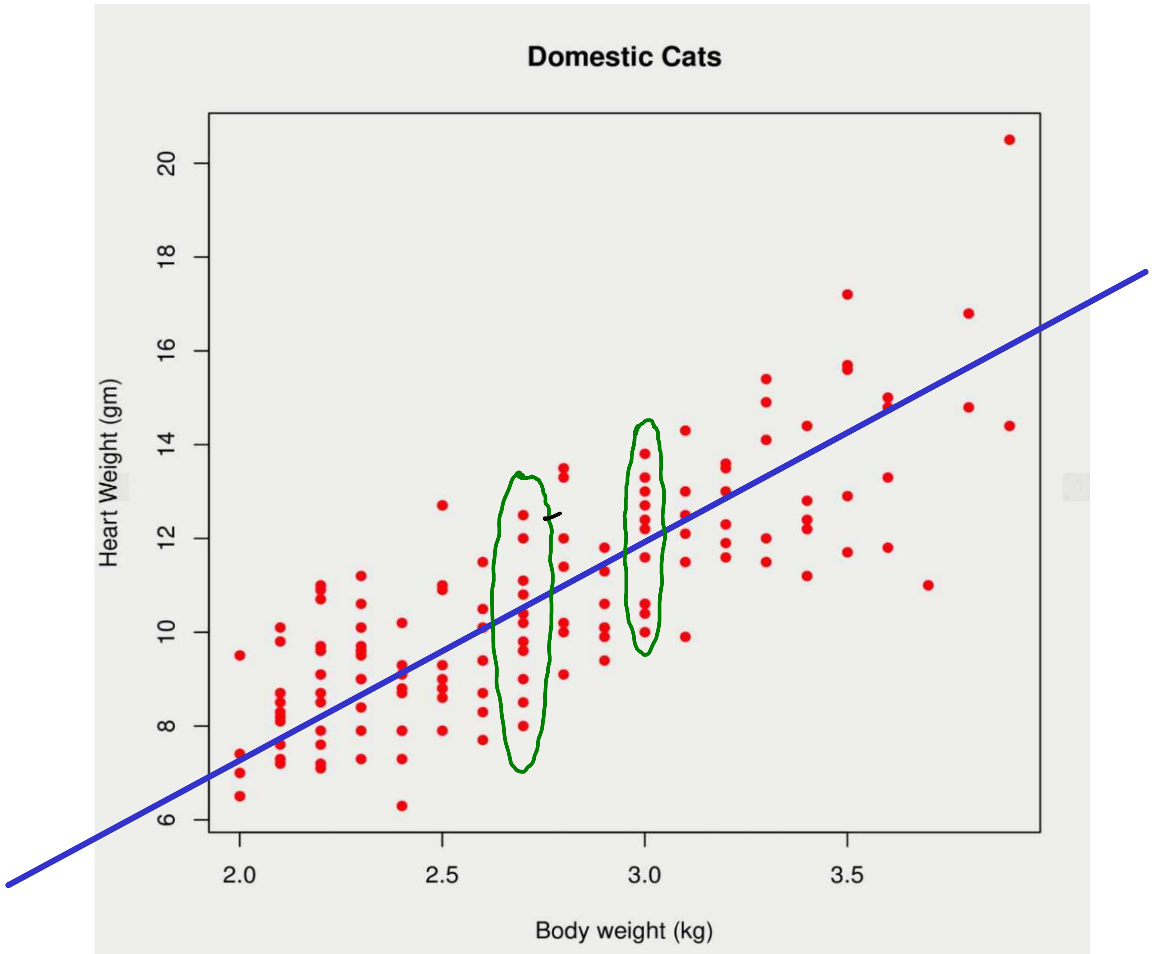
Outline

- Probabilistic models
- Some probabilistic concepts
  - Random variable, density function, normal distribution, joint density function, empirical distribution
- Maximum likelihood
  - Likelihood function and Maximum likelihood estimate
  - Learning via log-likelihood
  - Example: linear regression revisited

Cartoon picture of ML



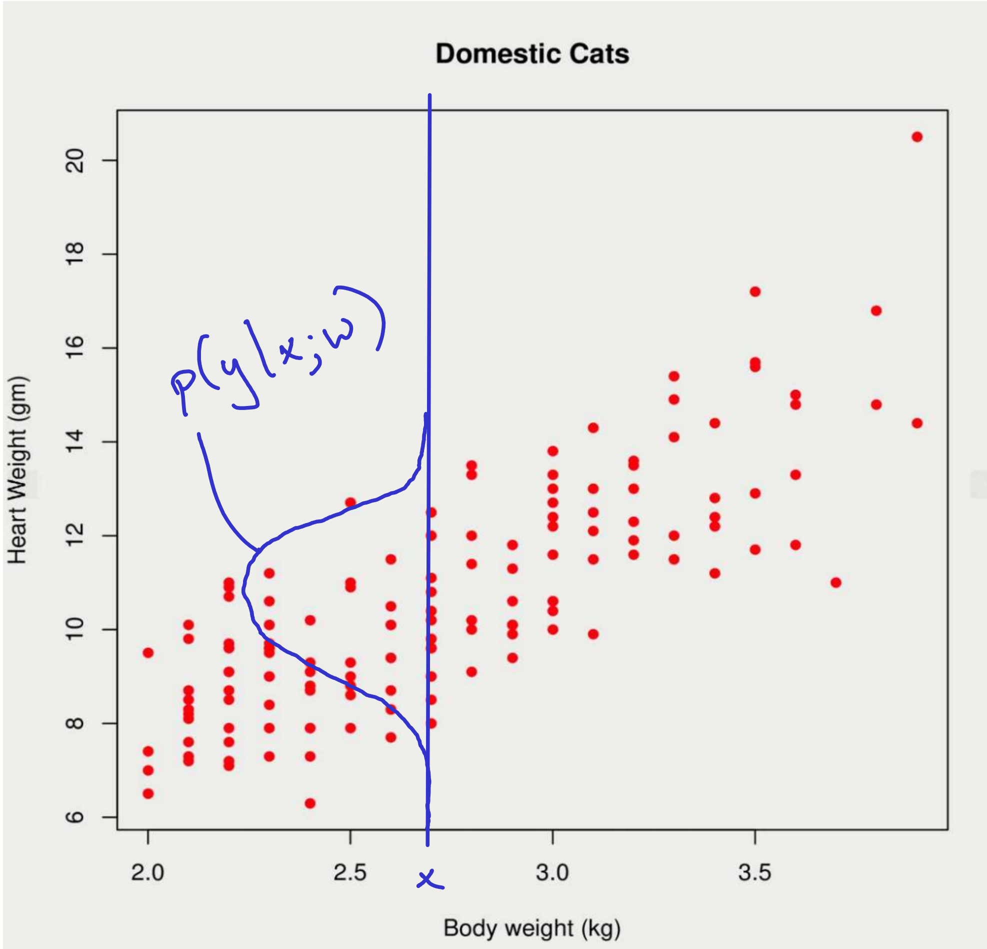
Prediction of Heart Weights Revisited



- We have considered a deterministic model

$$f(x) = wx,$$

- However there is variation in the data
  - there are cats with the same body weight, but different heart weights



- A probabilistic model can account for variance in the data, for example

$$F(x) = wx + N$$

where

$$N \sim \mathcal{N}(0, \sigma^2)$$

is a "noise term" which is normally distributed with expectation 0 and varience σ^2.

- $F(x)$ is a "random variable" which can be described by a "conditional density" $p(y|x;w)$

Random Variables and Probability Density Functions

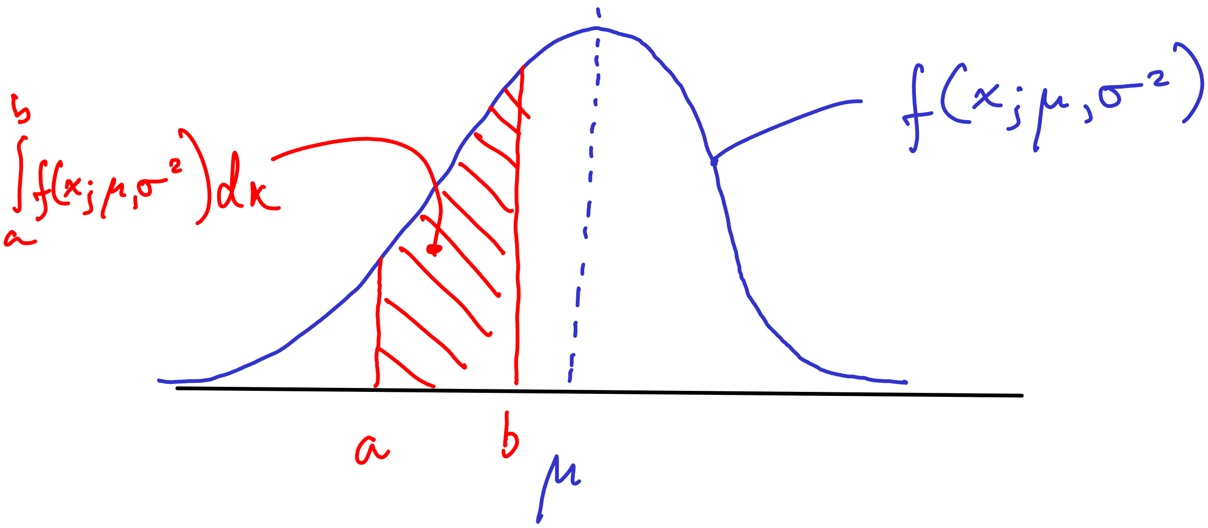
A random variable takes a value that depends on a random phenomenon, e.g.

- the number of dots when throwing a dice (6 possible values)
- the temperature at 9 am (infinitely many possible values)

The density function of a continuous random variable X is a function $p : \mathbb{R} \to \mathbb{R}$ s.t.

$$\int_a^b p(x)\, dx = \Pr(a \leq X \leq b)$$

Normal Distribution $N(\mu, \sigma^2)$



The normal distribution has probability density function

$$f(x;\underbrace{\mu,\sigma^2}_{\text{parameters of the distribution}}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Expectation

The expected value of $f(X)$ when X is a random variable with probability density function p is

$$\mathbb{E}_{X\sim p} f(X) := \int_{-\infty}^{\infty} p(x) f(x)\, dx$$

Example

$$\mathbb{E}_{X\sim N(\mu,\sigma^2)} X = \mu.$$

Joint Distributions and Independence

The joint density function of n random variables $X^{(1)}, \dots, X^{(n)}$ is a functon $p : \mathbb{R}^n \to \mathbb{R}$ such that

$$\int_D p\left(x^{(1)}, \dots, x^{(n)}\right) dx^{(1)} \cdots dx^{(n)} = \mathbb{P}_r\left(\left(X^{(1)}, \dots, X^{(n)}\right) \in D\right)$$

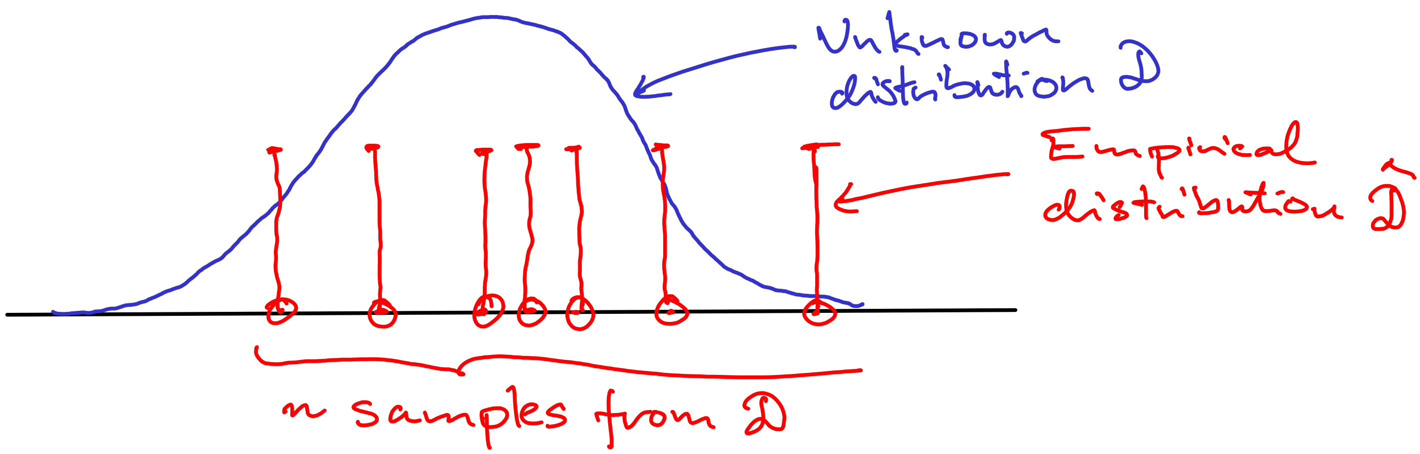
for any n-dimensional domain $D \subseteq \mathbb{R}^n$

If $X^{(1)}, \dots, X^{(n)}$ are n independent random variables with density functions $p^{(1)}, \dots, p^{(n)}$ and joint density p, then

$$p\left(x^{(1)}, \dots, x^{(n)}\right) = \prod_{i=1}^{n} p^{(i)}\left(x^{(i)}\right)$$

Empirical Distribution

Given n independent samples $X^{(1)}, \dots, X^{(n)}$ from an unknown distribution $\mathcal{D}$, we can construct an "approximation" of $\mathcal{D}$ by uniformly sampling from the set $\{X^{(1)}, \dots, X^{(n)}\}$.



Given $X^{(1)}, \dots, X^{(n)}$ i.i.d samples from $\mathcal{D}$, the empirical distribution of $\mathcal{D}$ has density function

$$\hat{P}_n(x) := \frac{1}{n} \sum_{i=1}^{n} \delta\left(X^{(i)} - x\right)$$

where δ is the Dirac delta, i.e.,

$\delta(x) = 0$ for $x \neq 0$ and $\int_{-\infty}^{\infty} \delta(x)\,dx = 1$.

NB $\mathbb{E}_{X \sim \hat{P}_n} f(X) = \frac{1}{n} \sum_{i=1}^{n} f(X^{(i)})$

Learning task revisited

Instead of deterministically predicting an output y for a given input x, we will train a probabilistic model represented by a conditional density function

density function of output

$$\underbrace{P_{model}}_{\text{density function of output}}(y \mid \underbrace{x}_{\text{input}} ; \underbrace{\theta}_{\text{parameter of model}})$$

Given training data and a family of probability models, we need to choose the parameter(s) θ which are appropriate for the data

$\Longrightarrow$ Maximum Likelihood Estimate

Likelihood function

Given independent training data $(x^{(1)}, y^{(1)}), \dots, (x^{(n)}, y^{(n)})$, and a probabilistic model P_{model} with parameter Θ, the <u>likelihood function</u> is defined as

$$\mathcal{L}(\underbrace{\Theta}_{\substack{\text{model}\\\text{parameter}\\\text{(variable)}}}; \overbrace{(x^{(1)}, y^{(1)}), \dots, (x^{(n)}, y^{(n)})}^{\text{training data (fixed)}}) := \prod_{i=1}^{n} \underbrace{P_{model}(y^{(i)} \mid x^{(i)}; \Theta)}_{\substack{\text{conditional}\\\text{density}}}$$

$\mathcal{L}(\Theta; \dots)$ is the "likelihood" that the observed data came from the model with parameter Θ.

Maximum Likelihood Estimate (MLE)

Given training data and a family of models indexed by a parameter θ, which of the models are most likely to have produced the data?

$$\theta_{MLE} := \underset{\theta}{\operatorname{argmax}} \mathcal{L}\left(\theta; (x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})\right)$$

$$= \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{n} P_{model}\left(y^{(i)} \mid x^{(i)}; \theta\right)$$

Log-likelihood

For numerical and analytical reasons, a convenient reformulation is

$$\Theta_{MLE} = \arg\max_{\Theta} \mathcal{L}(\Theta)$$ by def.

$$= \arg\max_{\Theta} \log \mathcal{L}(\theta)$$ log(x) monotonically increasing

$$= \arg\max_{\Theta} \log \prod_{i=1}^{n} P_{model}(y^{(i)} | x^{(i)}; \theta)$$

log(ab) = log a + log b

$$= \arg\max_{\theta} \sum_{i=1}^{n} \log P_{model}(y^{(i)} | x^{(i)}; \theta)$$

$$= \arg\min_{\Theta} \frac{1}{m} \sum - \log P_{model}(y^{(i)} | x^{(i)}; \theta)$$

Does not impact argmin

$$= \arg\min_{\Theta} \mathbb{E}_{(x,y)\sim\hat{D}} \left[-\log P_{model}(y | x; \theta) \right]$$

Learning via log-likelihood

Neural network models are often trained by minimising the negative log-likelihood of the model given the training data, i.e., by minimising the function

$$J(\theta) = \mathbb{E}_{(x,y)\sim\hat{D}}\left[-\log P_{model}(y|x;\theta)\right]$$

cost function

model parameter(s)

empirical distribution of data

Example: Linear Regression

Predicting heart weight from body weight with model

$$F(x) = wx + N \quad \text{where } N \sim N(0, \sigma^2),$$

hence

$$F(x) \sim N(wx, \sigma^2),$$

i.e., a model with conditional density

$$P_{model}(y \mid x; w) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y - xw)^2}{2\sigma^2}\right)$$

$$\Rightarrow w_{MLE} = \underset{w}{\operatorname{argmin}} \sum_{i=1}^{n} -\log P_{model}(y^{(i)} \mid x^{(i)}; w)$$

$$= \underset{w}{\operatorname{argmin}} \sum_{i=1}^{n} \left[\underbrace{-\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)}_{\text{constant}} + \frac{(y^{(i)} - x^{(i)}w)^2}{2\sigma^2} \right]$$

$$= \underset{w}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2}\left(y^{(i)} - x^{(i)}w\right)^2$$

$\Rightarrow$ Identical to MSE-solution!

Summary

- Probabilistic models
- Some probabilistic concepts
  - Random variable, density function, normal distribution, joint density function, empirical distribution
- Maximum likelihood
  - Likelihood function and Maximum likelihood estimate
  - Learning via log-likelihood. Example: linear regression
- Compulsory reading
  - Goodfellow et al., 3.1-3.8, 3.9.3, 5.5

Next lecture

- Gradient descent in higher dimensions