

# **Machine Learning and Data Mining**

## Recapitulation

---

Maxim Borisyak

Constructor University Bremen

February 16, 2026

# **Fundamentals of statistics**

---

# Two schools of statistics

---

## Frequentist statistics:

- probability = long-run frequency of events;
- parameters are fixed but unknown constants;
- inference based on sampling distributions;
- methods: MLE, hypothesis testing, confidence intervals.

## Bayesian statistics:

- probability = degree of belief;
- parameters are random variables with distributions;
- inference via Bayes' theorem:  $P(\theta | X) \propto P(X | \theta)P(\theta)$ ;
- methods: posterior distributions, credible intervals, MAP.

# Random variables

---

## Random variable:

- a measurable function  $X : \Omega \rightarrow \mathbb{R}$ ;
- maps outcomes from sample space to real numbers.

## Types:

- **discrete**:  $X \in \{x_1, x_2, \dots\}$ ;
  - probability mass function:  $P(X = x)$ ;
  - examples: coin flip, die roll, number of defects;
- **continuous**:  $X \in \mathbb{R}$ ;
  - probability density function:  $p(x)$ ;
  - examples: height, temperature, measurement error.

# **Random variables: discrete examples**

---

## **Discrete examples:**

- coin flip:  $X \in \{0, 1\}$ ,  $P(X = 1) = p$ ;
- die roll:  $X \in \{1, 2, 3, 4, 5, 6\}$ ,  $P(X = k) = 1/6$ ;
- number of customers per hour:  $X \in \{0, 1, 2, \dots\}$ .

## **Properties:**

- finite or countably infinite outcomes;
- $\sum_x P(X = x) = 1$ .

## Random variables: continuous examples

---

**Normal distribution:**  $X \sim \mathcal{N}(\mu, \sigma^2)$

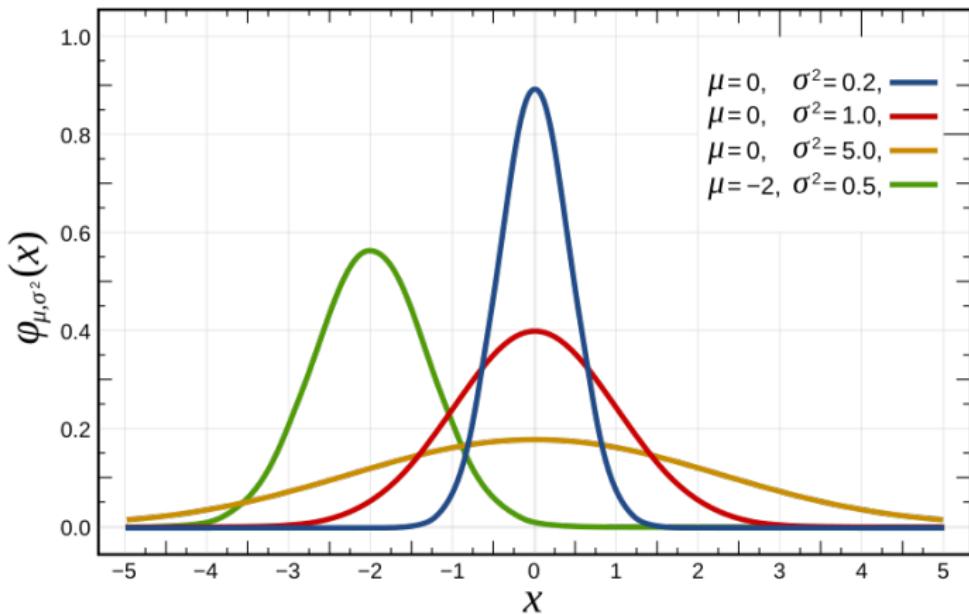
$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

**Examples:**

- person's height:  $X \sim \mathcal{N}(170, 10^2)$  cm;
- measurement error:  $X \sim \mathcal{N}(0, \sigma^2)$ .

## Normal distribution: visualization

---



Source: commons.wikimedia.org

# Multivariate normal distribution

---

**Multivariate normal:**  $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp \left[ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right]$$

where:

- $\mathbf{x} \in \mathbb{R}^d$  — random vector;
- $\boldsymbol{\mu} \in \mathbb{R}^d$  — mean vector:  $\boldsymbol{\mu}_i = \mathbb{E}[X_i]$ ;
- $\Sigma \in \mathbb{R}^{d \times d}$  — covariance matrix:  $\Sigma_{ij} = \text{Cov}[X_i, X_j]$ ;
- $|\Sigma|$  — determinant of  $\Sigma$ .

# Multivariate normal: covariance matrix

---

**Covariance matrix  $\Sigma$ :**

$$\Sigma_{ij} = \text{Cov}[X_i, X_j] = \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)]$$

**Properties:**

- symmetric:  $\Sigma = \Sigma^T$ ;
- positive semi-definite:  $\mathbf{v}^T \Sigma \mathbf{v} \geq 0$  for all  $\mathbf{v}$ ;
- diagonal elements:  $\Sigma_{ii} = \mathbb{D}[X_i]$ ;
- off-diagonal: correlation between variables.

## Multivariate normal: special cases

---

**Independent components:**  $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_d^2)$

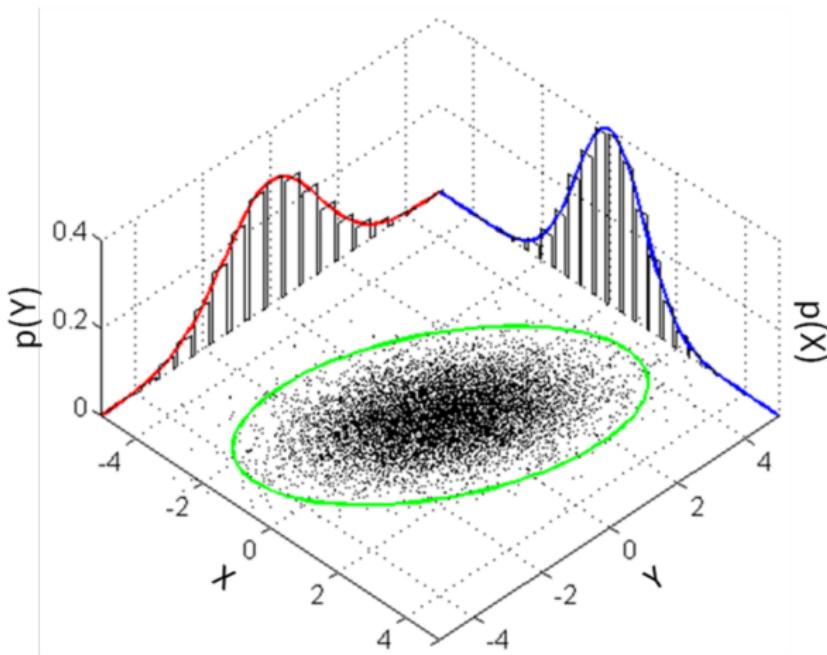
$$p(\mathbf{x}) = \prod_{i=1}^d \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{(x_i - \mu_i)^2}{2\sigma_i^2}\right]$$

**Spherical (isotropic):**  $\Sigma = \sigma^2 I$

$$p(\mathbf{x}) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left[-\frac{\|\mathbf{x} - \boldsymbol{\mu}\|^2}{2\sigma^2}\right]$$

## Multivariate normal: visualization

---



3D surface and contour plots of bivariate normal distribution.

Source: [commons.wikimedia.org](https://commons.wikimedia.org)

## Random variables: expectation

---

**Expectation** (mean):

$$\mathbb{E}[X] = \begin{cases} \sum_x x \cdot P(X = x) & \text{discrete} \\ \int x \cdot p(x) dx & \text{continuous} \end{cases}$$

**Variance:**

$$\begin{aligned}\mathbb{D}[X] &= \mathbb{E}[(X - \mathbb{E}[X])^2] \\ &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2\end{aligned}$$

# Statistical estimators

---

Given:

- sample:  $X = \{x_1, \dots, x_n\}$  from distribution  $P(x | \theta)$ ;
- unknown parameter:  $\theta$ .

**Estimator:**  $\hat{\theta}(X)$  — a function of sample.

**Examples:**

- sample mean:  $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i$ ;
- sample variance:  $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2$ .

# Bias of an estimator

---

**Bias:**

$$\text{Bias}[\hat{\theta}] = \mathbb{E}[\hat{\theta}] - \theta$$

**Unbiased estimator:**

$$\mathbb{E}[\hat{\theta}] = \theta \quad \Leftrightarrow \quad \text{Bias}[\hat{\theta}] = 0$$

## Asymptotically unbiased estimator

---

**Asymptotically unbiased:**

$$\lim_{n \rightarrow \infty} \mathbb{E}[\hat{\theta}_n] = \theta$$

or equivalently:

$$\lim_{n \rightarrow \infty} \text{Bias}[\hat{\theta}_n] = 0$$

- bias vanishes as sample size grows;
- weaker condition than unbiasedness.

## Examples: biased vs unbiased

---

**Sample mean** for  $\mu$ :

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i$$

- $\mathbb{E}[\hat{\mu}] = \mu$  — **unbiased**.

**Sample variance** for  $\sigma^2$ :

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2$$

- $\mathbb{E}[\hat{\sigma}^2] = \frac{n-1}{n} \sigma^2$  — **biased**;
- asymptotically unbiased:  $\lim_{n \rightarrow \infty} \mathbb{E}[\hat{\sigma}^2] = \sigma^2$ .

## Unbiased sample variance

---

**Unbiased estimator** for  $\sigma^2$ :

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \hat{\mu})^2$$

then:

$$\mathbb{E}[s^2] = \sigma^2$$

- division by  $n - 1$  (Bessel's correction) accounts for using estimated mean;
- loses one degree of freedom.

## Example: asymptotically unbiased wins

---

Compare variance estimators (assuming  $X_i \sim \mathcal{N}(\mu, \sigma^2)$ ):

**Unbiased:**  $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$

- $\mathbb{E}[s^2] = \sigma^2$ ;
- $\mathbb{D}[s^2] = \frac{2\sigma^4}{n-1}$ .

**MLE (biased):**  $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$

- $\mathbb{E}[\hat{\sigma}^2] = \frac{n-1}{n}\sigma^2$  — biased;
- $\mathbb{D}[\hat{\sigma}^2] = \frac{2(n-1)\sigma^4}{n^2} < \mathbb{D}[s^2]$  — **lower variance!**

## Example: asymptotically unbiased wins

---

### Comparison:

$$\frac{\mathbb{D}[\hat{\sigma}^2]}{\mathbb{D}[s^2]} = \frac{(n-1)^2}{n^2} \approx 1 - \frac{2}{n}$$

- MLE has  $\frac{2}{n}$  less variance;
- bias:  $\text{Bias}[\hat{\sigma}^2] = -\frac{\sigma^2}{n} \rightarrow 0$ ;
- for large  $n$ : MLE dominates in MSE;
- tradeoff: small bias buys significant variance reduction.

## Counter example: unbiased but bad

---

**Estimator using only first element:**

$$\hat{\mu}_1 = x_1$$

For estimating  $\mu = \mathbb{E}[X_i]$ :

- $\mathbb{E}[\hat{\mu}_1] = \mathbb{E}[x_1] = \mu$  — **unbiased**;
- $\mathbb{D}[\hat{\mu}_1] = \sigma^2$  — does **not** decrease with  $n$ ;
- **not consistent**: does not converge to  $\mu$ ;
- wasteful: ignores  $n - 1$  observations.

## Counter example: comparison

---

Compare  $\hat{\mu}_1 = x_1$  vs  $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i$ :

Property	$\hat{\mu}_1$	$\hat{\mu}$
Unbiased	✓	✓
Variance	$\sigma^2$	$\sigma^2/n$
Consistent	✗	✓
MSE	$\sigma^2$	$\sigma^2/n$

**Lesson:** unbiasedness alone is not sufficient for a good estimator.

# Mean Squared Error

---

## Mean Squared Error:

$$\begin{aligned}\text{MSE}[\hat{\theta}] &= \mathbb{E}[(\hat{\theta} - \theta)^2] \\ &= \text{Bias}[\hat{\theta}]^2 + \mathbb{D}[\hat{\theta}]\end{aligned}$$

## Bias-variance decomposition:

- unbiased estimator may have high variance;
- slightly biased estimator may have lower MSE;
- fundamental tradeoff in machine learning.

## Example: biased but better MSE

---

### Estimating variance with shrinkage:

Given  $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ , consider:

- unbiased:  $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ ;
- biased:  $\hat{\sigma}_c^2 = \frac{1}{n+1} \sum_{i=1}^n (x_i - \bar{x})^2$ .

For small  $n$ ,  $\text{MSE}[\hat{\sigma}_c^2] < \text{MSE}[s^2]$ !

**Tradeoff:** accepting bias reduces variance enough to lower MSE.

# Consistency

---

**Consistent estimator:**

$$\hat{\theta}_n \xrightarrow{P} \theta \quad \text{as } n \rightarrow \infty$$

i.e.,  $\forall \varepsilon > 0$ :

$$\lim_{n \rightarrow \infty} P(|\hat{\theta}_n - \theta| > \varepsilon) = 0$$

- converges in probability to true parameter;
- weaker than unbiasedness (concerns large  $n$  behavior).

## Example: inconsistent estimator

---

### Constant estimator:

$$\hat{\mu}_0 = c \quad (\text{constant, independent of data})$$

For estimating  $\mu = \mathbb{E}[X_i]$ :

- $\mathbb{E}[\hat{\mu}_0] = c$  — biased (unless  $c = \mu$  by chance);
- $\mathbb{D}[\hat{\mu}_0] = 0$  — zero variance!
- **not consistent:**  $\hat{\mu}_0 \not\rightarrow \mu$  as  $n \rightarrow \infty$ ;
- does not use data at all.

**Another example:** we already saw  $\hat{\mu}_1 = x_1$  is unbiased but inconsistent.

## Example: asymptotically unbiased but inconsistent

---

**Estimator with noise:**

$$\hat{\mu}_\varepsilon = \frac{1}{2} \left[ \frac{1}{N} \sum_{i=1}^N x_i + x_1 \right]. \quad (1)$$

**Properties:**

- $\mathbb{E}[\hat{\mu}_\varepsilon] = \mu$  — unbiased for all  $n$ ;
- $P(|\hat{\mu}_\varepsilon - \mu| > \varepsilon) \not\rightarrow 0$ .

## Example: consistent but biased

---

Shrinkage estimator:

$$\hat{\mu}_s = \frac{1}{N} \sum_{i=1}^N x_i + \frac{c}{n}$$

where  $c \neq 0$  is a constant.

Properties:

- $\mathbb{E}[\hat{\mu}_s] = \mu + \frac{c}{n}$  — biased for all finite  $n$ ;
- $\text{Bias}[\hat{\mu}_s] = \frac{c}{n} \rightarrow 0$  — asymptotically unbiased;
- $\hat{\mu}_s \xrightarrow{P} \mu$  — **consistent!**
- shows: consistency   unbiasedness for finite  $n$ .

## **Statistical estimations**

---

## Setup

---

Given:

- data:  $X = \{x_i\}_{i=1}^N$ ;
- parameterized family of distributions  $P(x | \theta)$ .

Problem:

- estimate  $\theta$ .

## Maximum likelihood estimation

---

$$\begin{aligned}L(\theta) &= P(X \mid \theta); \\ \hat{\theta} &= \arg \max_{\theta} L(\theta).\end{aligned}$$

$$\mathcal{L}(\theta) = -\log \prod_i P(x_i \mid \theta) = -\sum_i \log P(x_i \mid \theta)$$

- consistent estimation:  $\hat{\theta} \rightarrow \theta$  as  $N \rightarrow \infty$ ;
- *might be biased*;
- equal to MAP estimation with uniform prior.

## MLE: example

---

Given samples  $\{x_i\}_{i=1}^N$  from a normal distribution estimate its mean.

$$\mu = \arg \min_{\mu} \mathcal{L}(X) =$$

$$\arg \min_m u - \sum_i \log \left( \frac{1}{Z} \exp \left[ -\frac{(x_i - \mu)^2}{2\sigma^2} \right] \right) =$$

$$\arg \min_{\mu} \sum_i (x_i - \mu)^2 = \frac{1}{N} \sum_i x_i$$

# Properties of MLE

---

**Maximum Likelihood Estimator:**

$$\hat{\theta}_{\text{MLE}} = \arg \max_{\theta} P(X | \theta)$$

**Key properties** (under regularity conditions):

1. **Consistency:**  $\hat{\theta}_{\text{MLE}} \xrightarrow{P} \theta$  as  $n \rightarrow \infty$ ;
2. **Asymptotic normality:**  $\sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta) \xrightarrow{d} \mathcal{N}(0, I(\theta)^{-1})$ ;
3. **Asymptotic efficiency:** achieves Cramér-Rao lower bound.

## Properties of MLE: bias

---

**MLE is generally biased:**

- $\mathbb{E}[\hat{\theta}_{\text{MLE}}] \neq \theta$  for finite  $n$ ;
- **asymptotically unbiased:**  $\lim_{n \rightarrow \infty} \mathbb{E}[\hat{\theta}_{\text{MLE}}] = \theta$ ;
- bias often decreases as  $O(1/n)$ .

**Example:** MLE for variance  $\sigma^2$  in normal distribution:

$$\hat{\sigma}_{\text{MLE}}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{\mu})^2$$

is biased but asymptotically unbiased.

## Properties of MLE: efficiency

---

**Fisher Information:**

$$I(\theta) = \mathbb{E} \left[ \left( \frac{\partial \log P(X | \theta)}{\partial \theta} \right)^2 \right]$$

**Cramér-Rao bound:**

$$\mathbb{D}[\hat{\theta}] \geq \frac{1}{n \cdot I(\theta)}$$

**Asymptotically efficient:**

- MLE achieves this bound as  $n \rightarrow \infty$ ;
- no other consistent estimator has lower asymptotic variance.

# Properties of MLE: invariance

---

## Functional invariance:

If  $\hat{\theta}_{\text{MLE}}$  is MLE for  $\theta$ , then for any function  $g$ :

$$\widehat{g(\theta)}_{\text{MLE}} = g(\hat{\theta}_{\text{MLE}})$$

## Example:

- MLE for  $\mu$  in  $\mathcal{N}(\mu, \sigma^2)$ :  $\hat{\mu} = \bar{x}$ ;
- MLE for  $\mu^2$ :  $\widehat{\mu^2} = \bar{x}^2$  (not  $\overline{x^2}$ ).

## Bayesian inference

---

$$P(\theta \mid X) = \frac{1}{Z} P(X \mid \theta) P(\theta);$$

- often, posterior distribution of predictions is of the main interest:

$$P(f(x) = y \mid X) = \int \mathbb{I}[f(x, \theta) = y] P(\theta \mid X) d\theta$$

- with a few exceptions posterior is intractable;
- often, approximate inference is utilized instead.

## BI: example

---

Given samples  $\{x_i\}_{i=1}^N$  from a normal distribution estimate mean under a normal prior.

$$\begin{aligned} P(\mu \mid X) &= \frac{1}{Z} P(X \mid \mu) P(\mu) = \\ &\quad \frac{1}{Z} \exp \left[ -\frac{\mu^2}{2c^2} \right] \cdot \prod \exp \left[ -\frac{(x_i - \mu)^2}{2\sigma^2} \right] \end{aligned}$$

$$\log P(\mu \mid X) = -Z - \frac{\mu^2}{2c^2} - \sum_i \frac{(x_i - \mu)^2}{2\sigma^2}$$

## Maximum a posteriori estimation

---

$$\hat{\theta} = \arg \max_{\theta} P(\theta \mid X) = \arg \max_{\theta} P(X \mid \theta)P(\theta) =$$

$$\arg \min_{\theta} [-\log P(X \mid \theta) - \log P(\theta)] =$$

$$\arg \min_{\theta} [\text{neg log likelihood} + \text{penalty}]$$

$$\hat{\theta} = \arg \min_{\theta} \left[ -\log P(\theta) - \sum_i \log P(x_i \mid \theta) \right]$$

- sometimes called **structural loss**:
  - i.e. includes 'structure' of the predictor into the loss.

## MAP: example

---

Given samples  $\{x_i\}_{i=1}^N$  from a normal distribution estimate mean under a normal prior.

$$\hat{\mu} = \arg \max_{\mu} \log P(\mu \mid X) =$$

$$\arg \max_{\mu} \left[ -Z - \frac{\mu^2}{2c^2} - \sum_i \frac{(x_i - \mu)^2}{2\sigma^2} \right] =$$

$$\arg \min_{\mu} \left[ \lambda \mu^2 + \sum_i (x_i - \mu)^2 \right] = \frac{1}{N+\lambda} \sum_i x_i$$

# **Machine Learning**

---

# Structure of a Machine Learning problem

---

Given:

- description of the problem:
  - prior knowledge;
- data:
  - input space:  $\mathcal{X}$ ;
  - output space:  $\mathcal{Y}$ ;
- metric  $M$ .

Problem:

- find a learning algorithm:  $A : \mathcal{D} \rightarrow (\mathcal{X} \rightarrow \mathcal{Y})$  such that:

$$M(A(\text{data})) \rightarrow \max$$

## Differences from statistics

---

### Machine Learning:

- distributions are often intractable;
- high-dimensionality/small sample sizes;
- **universal approximators**;
- solves direct problem.

### Statistics:

- process modelling;
- low-dimensionality/large sample sizes;
- (approx.) prob. distributions;
- **exact inference**;
- infers process parameters.

# **Supervised learning**

---

# Regression

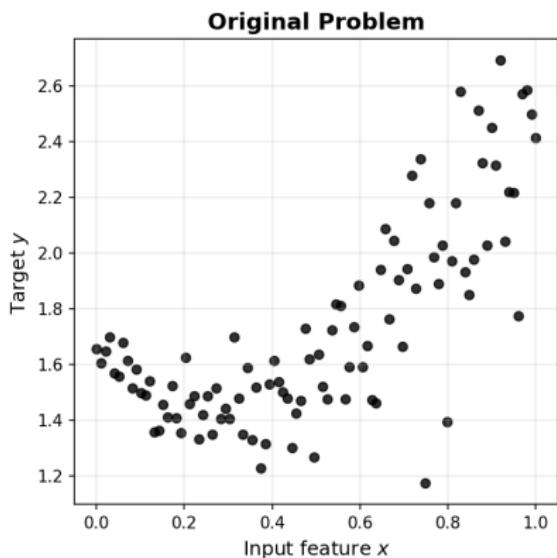
---

Input:  $x \in \mathbb{R}^n$ :

- samples;
- features;
- inputs;
- predictor (statistics).

Output:  $y \in \mathbb{R}^m$ :

- target;
- label;
- response.



# (Ordinary) Regression

Given a sample from:

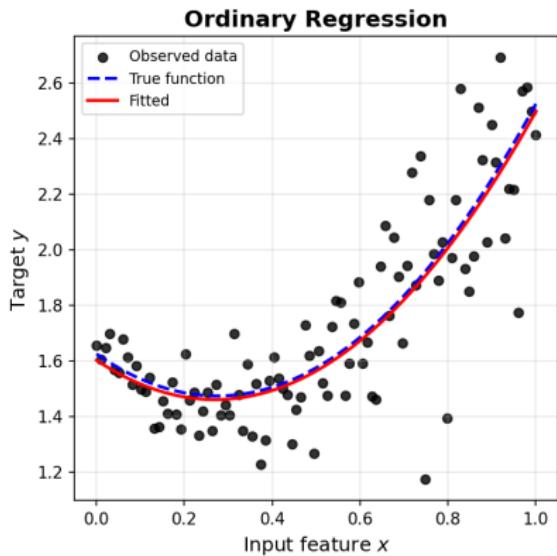
$$y = \hat{f}(x) + \varepsilon;$$

$$\varepsilon \sim P(\varepsilon | x);$$

$$\mathbb{E} [\varepsilon | x] = 0.$$

find a model  $m(x)$  such that:

$$m(x) \approx \mathbb{E} [y | x] = \hat{f}(x).$$



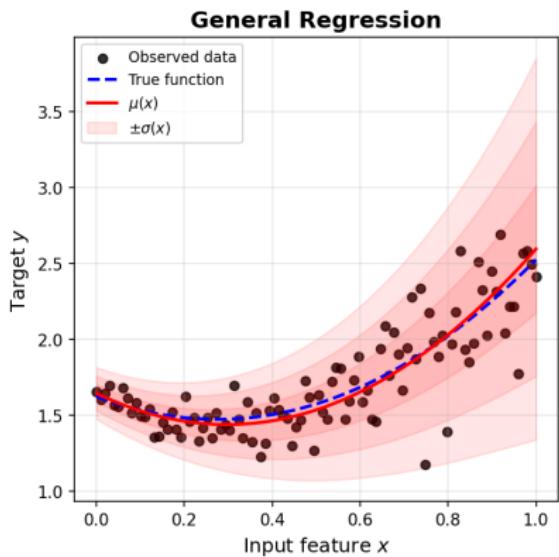
# General Regression

Given a sample from:

$$y = P(y | x);$$

find a model  $Q(y | x)$  such that:

$$Q(y | x) \approx P(y | x).$$



## Regression loss

---

$$\begin{aligned}\mathcal{L}(f) &= - \sum_i \log P_y(y_i \mid f, x_i) = \\ &\quad - \sum_i \log P_\varepsilon(y_i - f(x_i) \mid f, x_i) = \\ &\quad - \sum_i \log P_\varepsilon(y_i - f(x_i) \mid x_i)\end{aligned}$$

## Regression: MSE

---

- $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ ;
- $\sigma_\varepsilon^2 = \text{const}$  (unknown);

$$\begin{aligned}\mathcal{L}(f) &= -\sum_i \log P_\varepsilon(y_i - f(x_i) \mid x_i) = \\ &\sum_i \left[ Z(\sigma_\varepsilon^2) - \frac{(y_i - f(x_i))^2}{2\sigma_\varepsilon^2} \right] \sim \\ &\sum_i (y_i - f(x_i))^2 \rightarrow \min\end{aligned}$$

$$f^*(x) = \mathbb{E}[y \mid x]$$

## Regression: MSE

---

MSE always recovers the mean in the limit of infinite data.

Not always as efficient of MLE.

$$\begin{aligned}\mathcal{L}(f) &= \mathbb{E} \left[ (y - f(x))^2 \mid x \right] = \\ &\quad \mathbb{E} \left[ y^2 - 2yf(x) + f^2(x) \mid x \right] = \\ &\quad \mathbb{E} \left[ y_c^2 - 2y_c\mu(x) - 2y_c f(x) + f^2(x) - 2\mu(x)f(x) + \mu^2(x) \mid x \right] = \\ &\quad \sigma^2 + 0 + 0 + \mathbb{E} \left[ (f(x) - \mu(x))^2 \mid x \right]\end{aligned}$$

where:

- $y_c = y - \mathbb{E} y = y - \mu;$

## Regression: MAE

---

- $\varepsilon \sim \text{Laplace}(0, b_\varepsilon)$ ;
- $b_\varepsilon = \text{const}$  (unknown);

$$\begin{aligned}\mathcal{L}(f) &= - \sum_i \log P_\varepsilon(y_i - f(x_i) \mid x_i) = \\ &\quad \sum_i \left[ Z(b_\varepsilon) - \frac{|y_i - f(x_i)|}{2b_\varepsilon} \right] \sim \\ &\quad \sum_i |y_i - f(x_i)| \rightarrow \min\end{aligned}$$

$$f^*(x) = \text{median}[y \mid x]$$

# Linear regression

---

$$f(x) = w \cdot x$$

## Linear regression + MSE + MLE

---

$$\mathcal{L}(w) = \sum_i (w \cdot x_i - y_i)^2 = \|Xw - y\|^2 \rightarrow \min;$$

$$\frac{\partial}{\partial w} \mathcal{L}(w) = 2X^T(Xw - y) = 0;$$

$$w = (X^T X)^{-1} X^T y.$$

## Linear regression + MSE + MAP

---

$$\begin{aligned}\mathcal{L}(w) &= \sum_i (w \cdot x_i - y_i)^2 + \lambda \|w\|^2 = \\ &\quad \|Xw - y\|^2 + \lambda \|w\|^2 \rightarrow \min;\end{aligned}$$

$$\begin{aligned}\frac{\partial}{\partial w} \mathcal{L}(w) &= 2X^T(Xw - y) + 2\lambda w = 0; \\ w &= (X^T X + \lambda I)^{-1} X^T y.\end{aligned}$$

# Linear regression + MSE + Bayesian Inference

---

- prior:

$$w \sim \mathcal{N}(0, \Sigma_w);$$

- data model:

$$\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2).$$

# Linear regression + MSE + Bayesian Inference

---

$$P(w \mid y, X) \propto P(y \mid w, X)P(w) \propto$$

$$\exp \left[ -\frac{1}{2\sigma_\varepsilon^2} (y - Xw)^T (y - Xw) \right] \cdot \exp \left[ -\frac{1}{2} w^T \Sigma_w^{-1} w \right] =$$

$$\exp \left[ -\frac{1}{2} (w - w^*)^T A_w (w - w^*) \right]$$

where:

- $A_w = \frac{1}{\sigma_\varepsilon^2} XX^T + \Sigma_w^{-1};$
- $w^* = \frac{1}{\sigma_\varepsilon^2} A_w^{-1} Xy.$

## Linear regression + MSE + Bayesian Inference

---

To make prediction  $y'$  in point  $x'$ :

$$\begin{aligned} P(y' \mid y, X, x') &= \\ \int P(y' \mid w, x') P(w \mid X, y) &= \\ \mathcal{N}\left(\frac{1}{\sigma_\varepsilon^2} x'^T A^{-1} X y, x'^T A^{-1} x'\right) \end{aligned}$$

## Basis expansion

---

To capture more complex dependencies basis functions can be introduced:

$$f(x) = \sum_i w \cdot \phi(x)$$

where:

- $\phi(x) \in \mathbb{R}^K$  — expanded basis.
- $\phi$  **is fixed**.

## Basis expansion: example

---

Regression with polynomials:

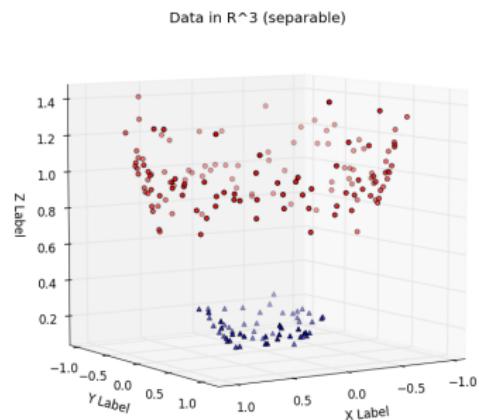
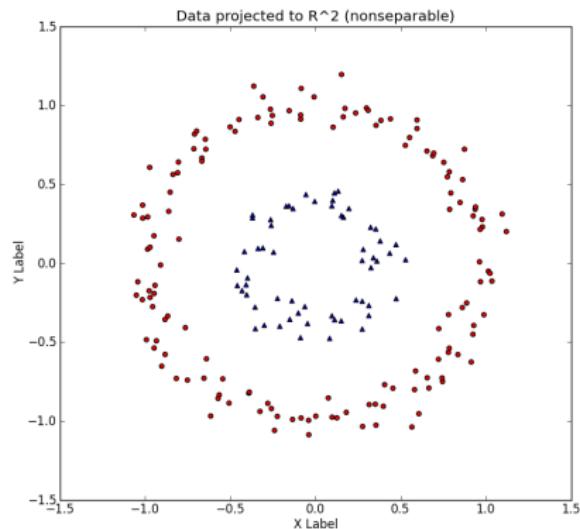
$$\phi(x) = \{1, x_1, \dots, x_n, x_1^2, x_1 x_2, \dots, x_n^2, \dots\}$$

Periodic functions:

$$\phi(x) = \{1, \sin(x), \cos(x), \sin(2x), \cos(2x), \dots\}$$

## Basis expansion: example

---



Source: eric-kim.net

# Classification

---

- classes:  $y \in \{1, 2, \dots, m\}$ ;
- classifier:

$$f: \mathcal{X} \rightarrow \mathbb{R}^m; \\ \sum_{k=1}^m f^k(x) = 1.$$

$$\mathcal{L}(f) = - \sum_i \sum_{k=1}^m \mathbb{I}[y_i = k] \log f^k(x_i); \\ \text{cross-entropy}(f) = \sum_i y'_i \cdot f(x_i).$$

# Softmax

---

- often employed trick to make  $f(x)$  a proper distribution:

$$f(x) = \text{softmax}(g(x));$$

$$f^i(x) = \frac{\exp(g^i(x))}{\sum_k \exp(g^k(x))}.$$

# Logistic regression

---

$$\begin{aligned}g(x) &= Wx + b; \\f(x) &= \text{softmax}(g(x)).\end{aligned}$$

Another form:

$$\frac{\log P(y = i \mid x)}{\log P(y = j \mid x)} = \frac{w_i \cdot x + b_i}{w_j \cdot x + b_j}$$

## Logistic regression: 2 classes

---

$$f_1(x) = \frac{\exp(w_1 \cdot x + b_1)}{\exp(w_1 \cdot x + b_1) + \exp(w_2 \cdot x + b_2)} =$$

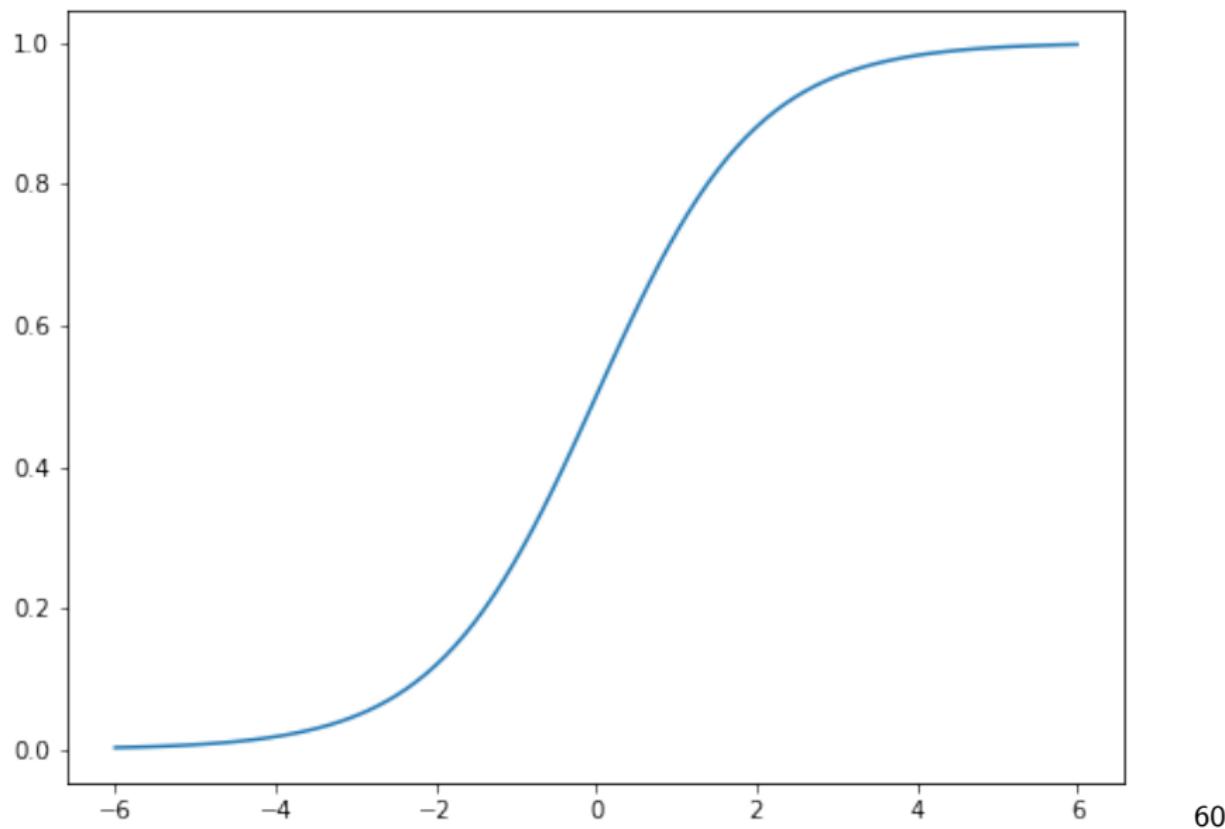
$$\frac{1}{1 + \exp((w_2 - w_1) \cdot x + b_2 - b_1)} =$$

$$\frac{1}{1 + \exp(w' \cdot x + b')} =$$

sigmoid( $w' \cdot x + b'$ ).

## Logistic regression: 2 classes

---



## Training logistic regression

---

$$\mathcal{L}(w) = \sum_i \mathbb{I}[y_i = 1] \log(1 + \exp(wx_i + b)) + \mathbb{I}[y_i = 0] \log(1 + \exp(-wx_i - b))$$

- has no analytical solution;
- smooth and convex.

## Gradient Descent

---

$$f(\theta) \rightarrow \min;$$

$$\theta^* = \arg \min_{\theta} f(\theta).$$

$$\theta^{t+1} = \theta^t - \alpha \nabla f(\theta^t);$$

$$\theta^t \rightarrow \theta^*, t \rightarrow \infty;$$

# Gradient Descent

---

```
1:  $\theta :=$  initialization  
2: for  $t := 1, \dots$  do  
3:    $\theta := \theta - \alpha \nabla f(\theta^t)$   
4: end for
```

# Stochastic Gradient Descent

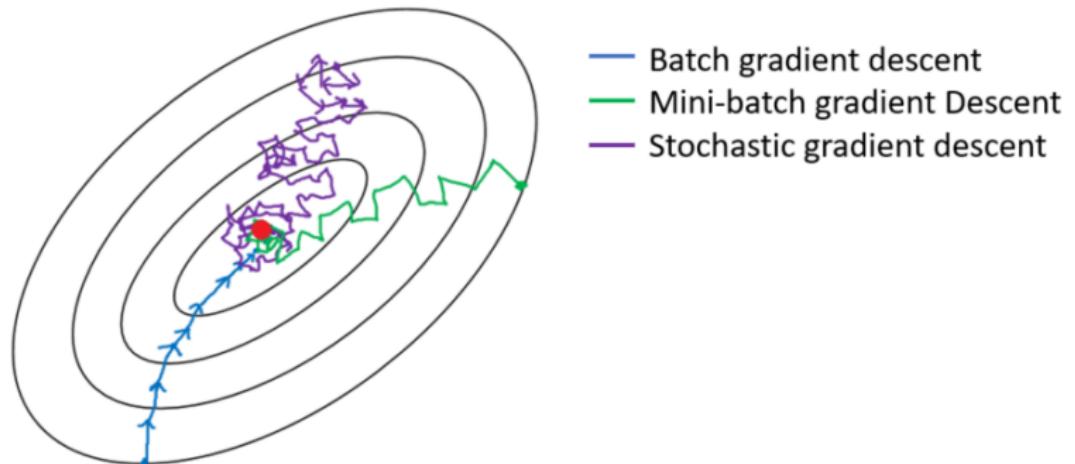
---

$$f(\theta) = \sum_{i=1}^N f_i(\theta)$$

```
1:  $\theta :=$  initialization  
2: for  $t := 1, \dots$  do  
3:    $i := \text{random}(1, \dots, N)$   
4:    $\theta := \theta - \alpha \nabla f_i(\theta^t)$   
5: end for
```

## Illustration

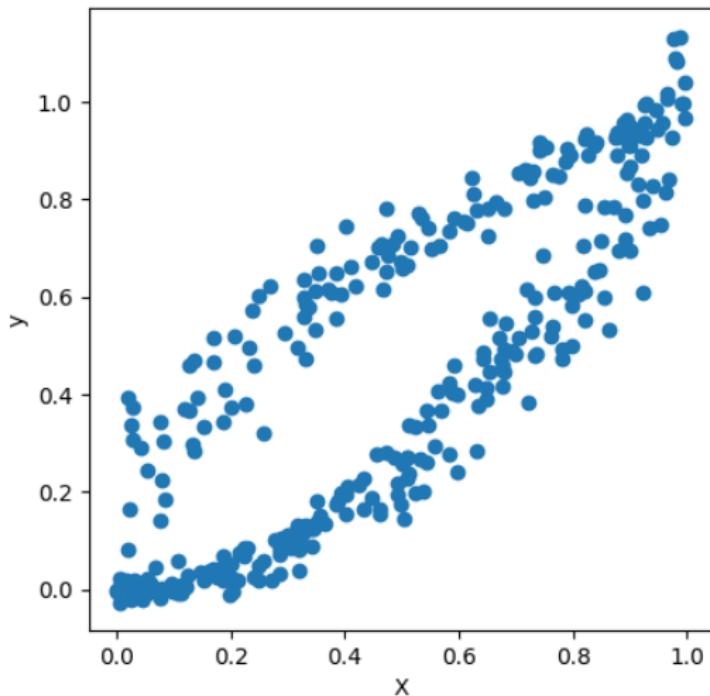
---



Source: [towardsdatascience.com](https://towardsdatascience.com/)

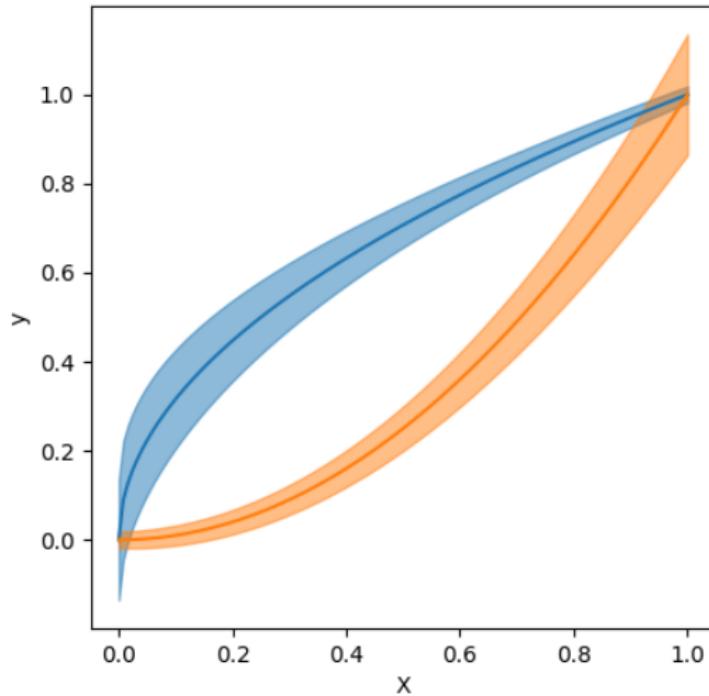
## Tricky example

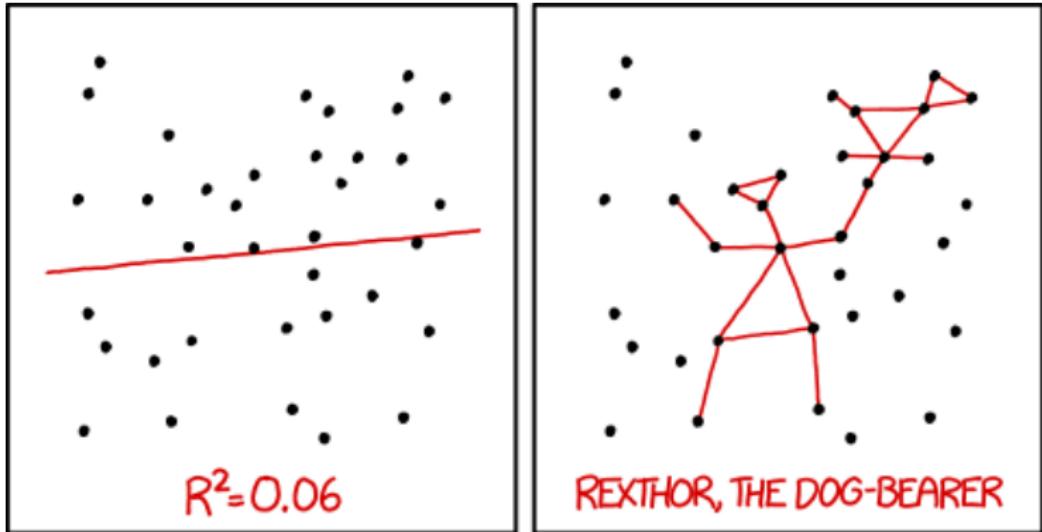
---



## Tricky example

---





I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER  
TO GUESS THE DIRECTION OF THE CORRELATION FROM THE  
SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

# **My first neural network**

---

# Universal Approximators

---

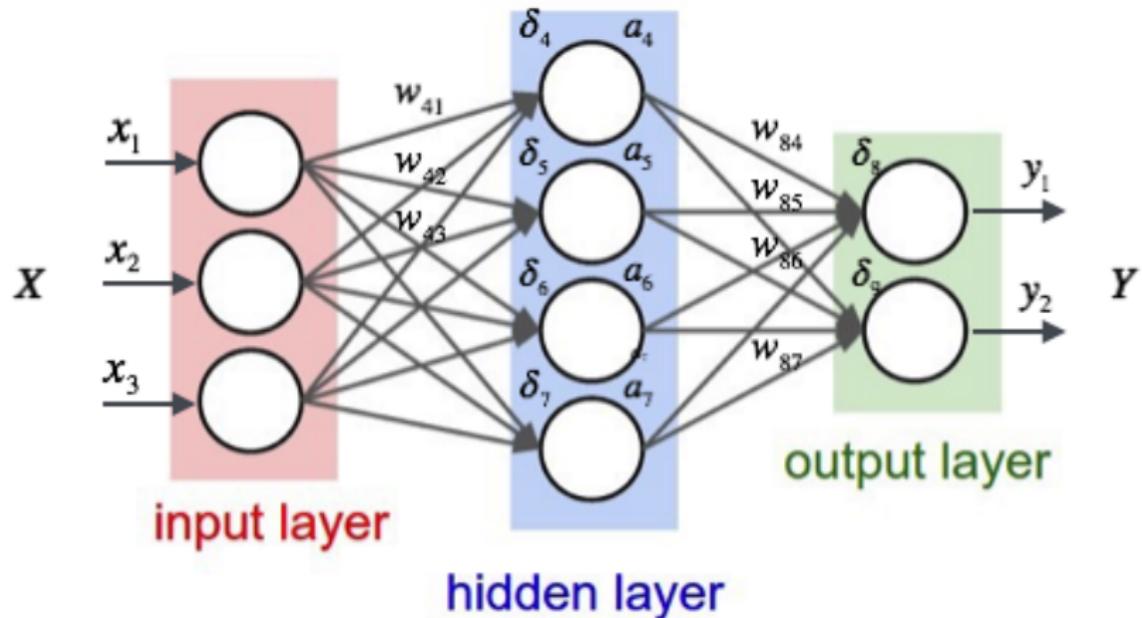
## Universal Approximation Theorem

If  $\phi$  is a non-constant, continuous, bounded, monotonic function, then every continuous function  $f$  on a compact set from  $\mathbb{R}^n$  can be approximated with any precision  $\varepsilon > 0$  by:

$$g(x) = c + \sum_{i=1}^N \alpha_i \phi(w_i \cdot x + b_i)$$

given large enough  $N$ .

# Universal Approximators



# How to train a neural network

---

Stochastic Gradient Descent and Co.

# How to train a neural network

---

Stochastic Gradient Descent and Co.

- how to initialize?
- how to choose an appropriate learning rate?
- how many units?
- which activation function to choose?