Machine Learning and Data Mining

Recapitulation

Maxim Borisyak

Constructor University Bremen

February 23, 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: $P(\theta \mid X) \propto P(X \mid \theta)P(\theta)$;
- methods: posterior distributions, credible intervals, MAP.

Random variables

Random variable:

- a measurable function $X : \Omega \to \mathbb{R}$;
- 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

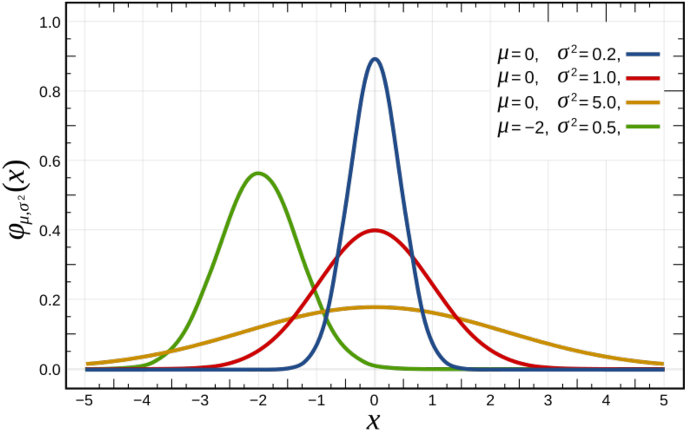
Normal distribution: $X \sim \mathcal{N}(\mu, \sigma^2)$

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

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: $\Sigma_{ij} = \mathrm{Cov}[X_i, X_j]$;
- $|\Sigma|$ — determinant of Σ.

Multivariate normal: covariance matrix

Covariance matrix Σ:

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

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

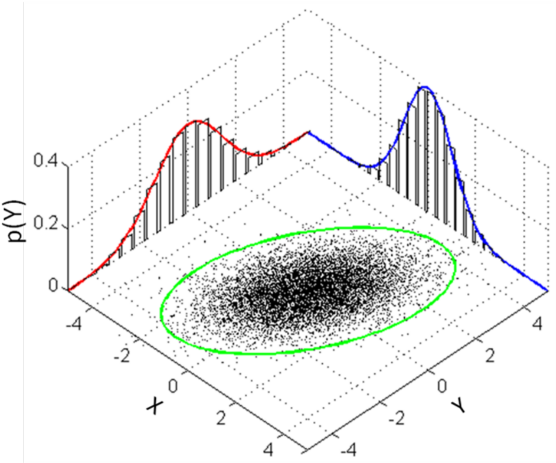
Independent components: $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$

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

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`

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:

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

Statistical estimators

Given:

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

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:

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

Unbiased estimator:

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

Asymptotically unbiased estimator

Asymptotically unbiased:

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

or equivalently:

$$\lim_{n\to\infty} \mathrm{Bias}[\hat{\theta}_n] = 0$$

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

Examples: biased vs unbiased

Sample mean for μ:

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

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

Sample variance for σ^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: $\lim_{n\to\infty} \mathbb{E}[\hat{\sigma}^2] = \sigma^2$.

Unbiased sample variance

Unbiased estimator for σ^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: $\mathrm{Bias}[\hat{\sigma}^2] = -\frac{\sigma^2}{n} \to 0$;
- 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 μ;
- 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	σ^2	σ^2/n
Consistent	$\times$	✓
MSE	σ^2	σ^2/n

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

Mean Squared Error

Mean Squared Error:

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

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 $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, 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, $\mathrm{MSE}[\hat{\sigma}_c^2] < \mathrm{MSE}[s^2]$!

Tradeoff: accepting bias reduces variance enough to lower MSE.

Consistency

Consistent estimator:

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

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

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

- 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**: $\hat{\mu}_0 \not\to \mu$ as $n \to \infty$;
- 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:

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

Properties:

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

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;
- $\mathrm{Bias}[\hat{\mu}_s] = \frac{c}{n} \to 0$ — 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 $P(x \mid \theta)$.

Problem:

- estimate θ.

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: $\hat{\theta} \to \theta$ as $N \to \infty$;
- *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:

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

Key properties (under regularity conditions):

1. **Consistency**: $\hat{\theta}_{\text{MLE}} \xrightarrow{P} \theta$ as $n \to \infty$;
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**: $\lim_{n\to\infty} \mathbb{E}[\hat{\theta}_{\text{MLE}}] = \theta$;
- bias often decreases as $O(1/n)$.

Example: MLE for variance σ^2 in normal distribution:

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

is biased but asymptotically unbiased.

Properties of MLE: efficiency

Fisher Information:

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

Cramér-Rao bound:

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

Asymptotically efficient:

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

Properties of MLE: invariance

Functional invariance:

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

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

Example:

- MLE for μ in $\mathcal{N}(\mu, \sigma^2)$: $\hat{\mu} = \bar{x}$;
- MLE for μ^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:

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

- 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.

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

$$\frac{1}{Z} \exp\left[-\frac{\mu^2}{2c^2}\right] \cdot \prod \exp\left[-\frac{(x_i - \mu)^2}{2\sigma^2}\right]$$

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

Maximum a posteriori estimation

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

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

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

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

- 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: $A : \mathcal{D} \to (\mathcal{X} \to \mathcal{Y})$ such that:

$$M(A(\mathrm{data})) \to \max$$

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

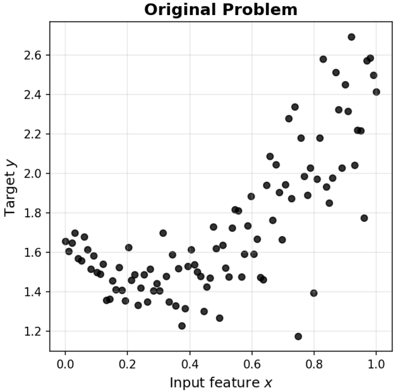
Regression

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:

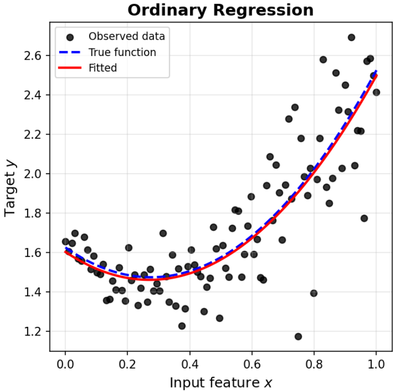
$$y = \hat{f}(x) + \varepsilon;$$

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

$$\mathbb{E}\left[\varepsilon \mid x\right] = 0.$$

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

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



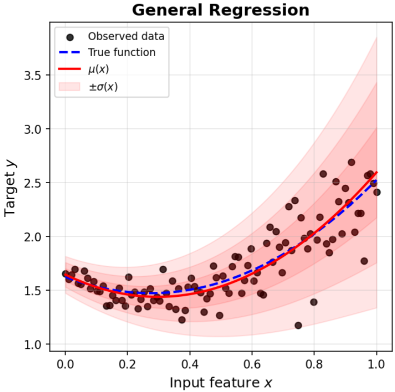
General Regression

Given a sample from:

$$y = P(y \mid x);$$

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

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



Regression loss

$$\mathcal{L}(f) = -\sum_i \log P_y(y_i \mid f, x_i) =$$

$$-\sum_i \log P_\varepsilon(y_i - f(x_i) \mid f, x_i) =$$

$$-\sum_i \log P_\varepsilon(y_i - f(x_i) \mid x_i)$$

Regression: MSE

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

$$\mathcal{L}(f) = -\sum_i \log P_\varepsilon(y_i - f(x_i) \mid x_i) =$$

$$\sum_i \left[Z(\sigma_\varepsilon^2) - \frac{(y_i - f(x_i))^2}{2\sigma_\varepsilon^2} \right] \sim$$

$$\sum_i (y_i - f(x_i))^2 \to \min$$

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

Regression: MSE

MSE always recovers the mean in the limit of infinite data.
Not always as efficient of MLE.

$$\mathcal{L}(f) = \mathbb{E}\left[(y - f(x))^2 \;\middle|\; x\right] =$$

$$\mathbb{E}\left[y^2 - 2y f(x) + f^2(x) \;\middle|\; x\right] =$$

$$\mathbb{E}\left[y_c^2 - 2y_c\mu(x) - 2y_c f(x) + f^2(x) - 2\mu(x)f(x) + \mu^2(x) \;\middle|\; x\right] =$$

$$\sigma^2 + 0 + 0 + \mathbb{E}\left[(f(x) - \mu(x))^2 \;\middle|\; x\right]$$

where:

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

Regression: MAE

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

$$\mathcal{L}(f) = -\sum_i \log P_\varepsilon(y_i - f(x_i) \mid x_i) =$$

$$\sum_i \left[Z(b_\varepsilon) - \frac{|y_i - f(x_i)|}{2 b_\varepsilon} \right] \sim$$

$$\sum_i |y_i - f(x_i)| \to \min$$

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

Linear regression

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

Linear regression + MSE + MLE

$$\begin{aligned}
\mathcal{L}(w) &= \sum_i (w \cdot x_i - y_i)^2 = \|Xw - y\|^2 \to \min; \\
\frac{\partial}{\partial w}\mathcal{L}(w) &= 2X^T(Xw - y) = 0; \\
w &= (X^T X)^{-1} X^T y.
\end{aligned}$$

Linear regression + MSE + MAP

$$\begin{aligned}
\mathcal{L}(w) &= \sum_i (w \cdot x_i - y_i)^2 + \lambda \|w\|^2 = \\
&\quad \|Xw - y\|^2 + \lambda \|w\|^2 \to \min; \\
\frac{\partial}{\partial w}\mathcal{L}(w) &= 2X^T(Xw - y) + 2\lambda w = 0; \\
w &= (X^T X + \lambda I)^{-1} X^T y.
\end{aligned}$$

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:

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

Linear regression + MSE + Bayesian Inference

To make prediction y' in point x':

$$P(y' \mid y, X, x') =$$

$$\int P(y' \mid w, x') P(w \mid X, y) =$$

$$\mathcal{N}\left(\frac{1}{\sigma_\varepsilon^2} x'^T A^{-1} X y, x'^T A^{-1} x'\right)$$

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.
- ϕ **is fixed**.

Basis expansion: example

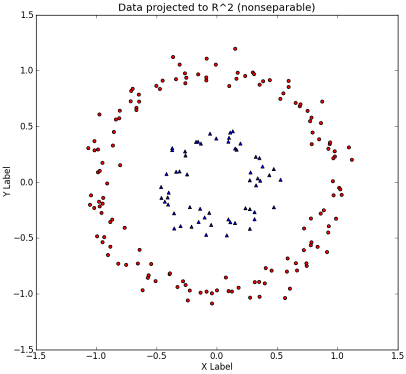
Regression with polynomials:

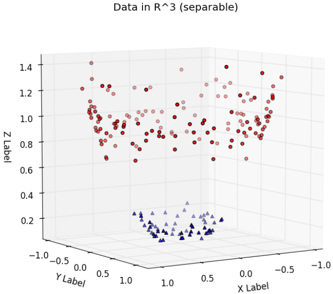
$$\phi(x) = \{1, x_1, \dots, x_n, x_1^2, x_1 x_2, \dots, x_n^2, \dots\}$$

Periodic functions:

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

Basis expansion: example





Source: `eric-kim.net`

Kernel methods

Kernel methods: motivation

Basis expansion $f(x) = w \cdot \phi(x)$ is powerful but costly:

- polynomial features of degree d on $\mathbb{R}^n$: $O(n^d)$ features;
- to capture complex structure, we may need $K \to \infty$;
- training in primal space costs $O(K^3)$.

Kernel methods

Theorem

The optimal w^* always lies in the span of training features:

$$w^* = \sum_{i=1}^{n} \alpha_i \phi(x_i)$$

$$f^*(x) = x \cdot w^* = x \cdot \left(\sum_{i=1}^{n} \alpha_i \, \phi(x_i) \right) = \sum_i \alpha_i \, (x \cdot x_i) \, .$$

$\Rightarrow$ we can work with **scalar products** only: $\phi(x) \cdot \phi(x')$.

Primal: ridge regression in feature space

Given feature map $\phi : \mathcal{X} \to \mathbb{R}^K$:

$$\mathcal{L}(w) = \|\Phi w - y\|^2 + \lambda \|w\|^2 \to \min_w$$

where $\Phi \in \mathbb{R}^{n \times K}$, $\Phi_{ij} = \phi_j(x_i)$.

Primal solution:

$$\underbrace{w^*}_{K} = \left(\underbrace{\Phi^T \Phi + \lambda I_K}_{K \times K} \right)^{-1} \underbrace{\Phi^T}_{K \times n} \underbrace{y}_{n}$$

From primal to dual

The optimality condition $\nabla_w \mathcal{L} = 0$ gives:

$$\begin{aligned} 2\Phi^T(\Phi w - y) + 2\lambda w &= 0 \\ w &= \frac{1}{\lambda}\Phi^T \underbrace{(y - \Phi w)}_{=:\lambda\alpha} = \Phi^T \alpha \end{aligned}$$

Substituting $w = \Phi^T \alpha$ back:

$$\begin{aligned} \Phi\Phi^T\alpha + \lambda\alpha &= y \\ (K + \lambda I_n)\alpha &= y \end{aligned}$$

where $K_{ij} = \phi(x_i) \cdot \phi(x_j)$ is the **Gram matrix**.

Theorem: primal–dual equivalence

Theorem

The primal and dual solutions are equivalent:

$$w^* = (\Phi^T \Phi + \lambda I_K)^{-1} \Phi^T y;$$

$$\alpha^* = (K + \lambda I_n)^{-1} y.$$

$$w^* = \Phi^T \alpha^*$$

Two forms of solution:

$$f_p(x) = w \cdot \phi(x);$$
$$f_k(x) = \sum_i \alpha_i \, \phi(x) \cdot \phi(x_i).$$

Kernel trick

Define the **kernel function** $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$:

$$k(x, x') = \phi(x) \cdot \phi(x').$$

The dual solution only depends on k:

$$\alpha^* = (K + \lambda I)^{-1} y, \qquad K_{ij} = k(x_i, x_j),$$
$$f(x') = \sum_i \alpha_i^* \, k(x_i, x').$$

Kernel

Definition. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a **kernel** if it is:

- **symmetric**: $k(x, x') = k(x', x)$;
- **positive semi-definite** (PSD): for all n, all $x_1, \ldots, x_n \in \mathcal{X}$, $c_1, \ldots, c_n \in \mathbb{R}$:

$$\sum_{i,j} c_i \, c_j \, k(x_i, x_j) \geq 0$$

Equivalently, the Gram matrix $K_{ij} = k(x_i, x_j)$ is PSD for any finite set of points.

Feature maps induce kernels

Every feature map $\phi : \mathcal{X} \to \mathcal{H}$ into a Hilbert space induces a kernel:

$$k(x, x') = \phi(x) \cdot \phi(x')$$

- **symmetry**: $\phi(x) \cdot \phi(x') = \phi(x') \cdot \phi(x)$;
- **PSD**: $\sum_{i,j} c_i c_j \, \phi(x_i) \cdot \phi(x_j) = \|\sum_i c_i \, \phi(x_i)\|^2 \geq 0$.

Every kernel is a scalar product

Theorem

Theorem (Mercer). *$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel if and only if there exists a Hilbert space $\mathcal{H}$ and a feature map $\phi : \mathcal{X} \to \mathcal{H}$ such that:*

$$k(x, x') = \phi(x) \cdot \phi(x').$$

- the feature space $\mathcal{H}$ may be infinite-dimensional;
- it is not unique — many $(\mathcal{H}, \phi)$ can realise the same k;
- the canonical choice is the RKHS $\mathcal{H}_k$ with $\phi(x) = k(x, \cdot)$.

Representer theorem

Theorem (A Generalized Representer Theorem)

Let $\mathcal{H}_k$ be an RKHS and consider:

$$\min_{f \in \mathcal{H}_k} \sum_{i=1}^{n} L(y_i, f(x_i)) + \Omega(\|f\|_{\mathcal{H}_k})$$

where L is any loss and $\Omega : \mathbb{R}_{\geq 0} \to \mathbb{R}$ is strictly monotone increasing.

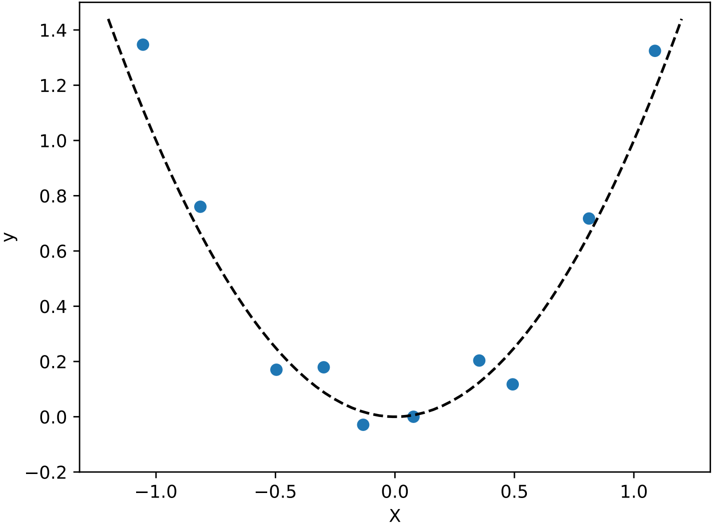
Then every minimizer admits the representation:

$$f^*(x) = \sum_{i=1}^{n} \alpha_i \, k(x_i, x)$$

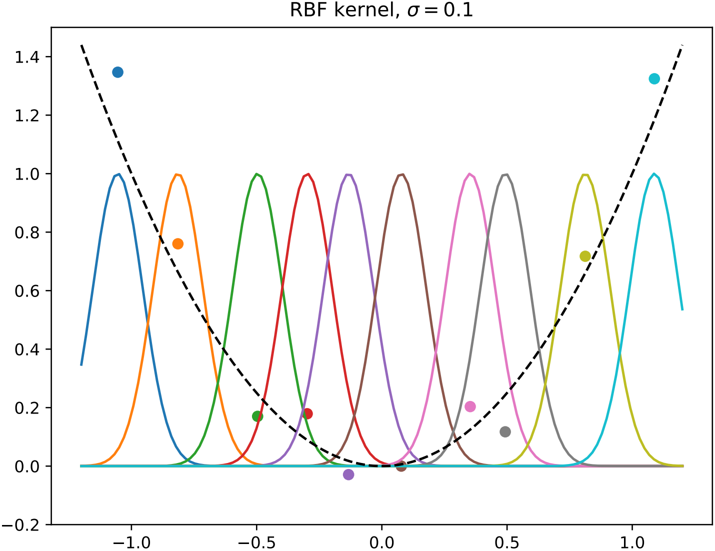
Common kernels

Kernel	$k(x, x')$	**Notes**
Linear	$x \cdot x'$	$\phi(x) = x$
Polynomial	$(x \cdot x' + c)^d$	degree-d features
RBF / Gaussian	$\exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right)$	∞-dim ϕ
Laplace	$\exp\left(-\frac{\|x - x'\|}{\sigma}\right)$	less smooth than RBF
Matérn	(various)	controlled smoothness

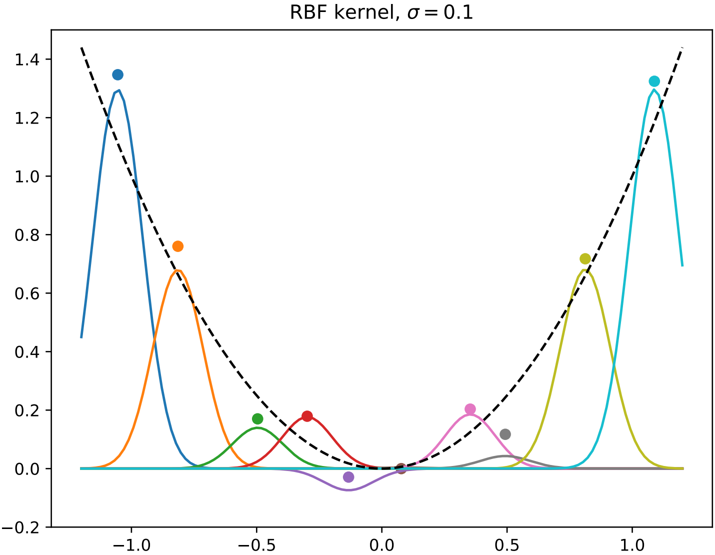
Example



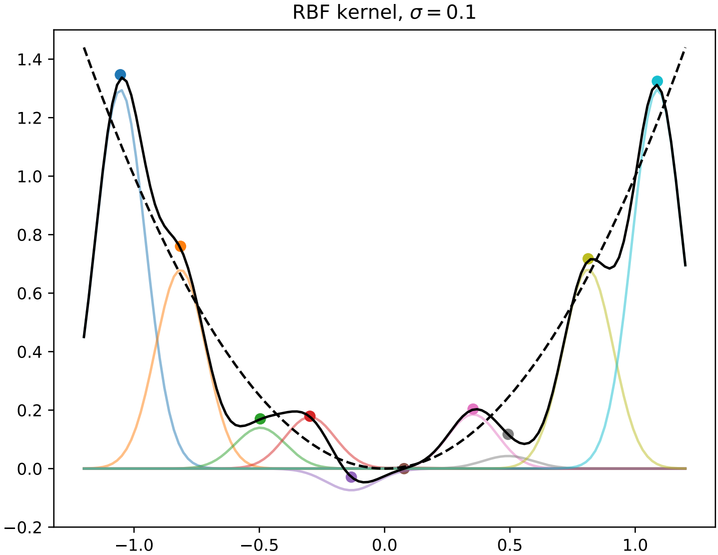
Example



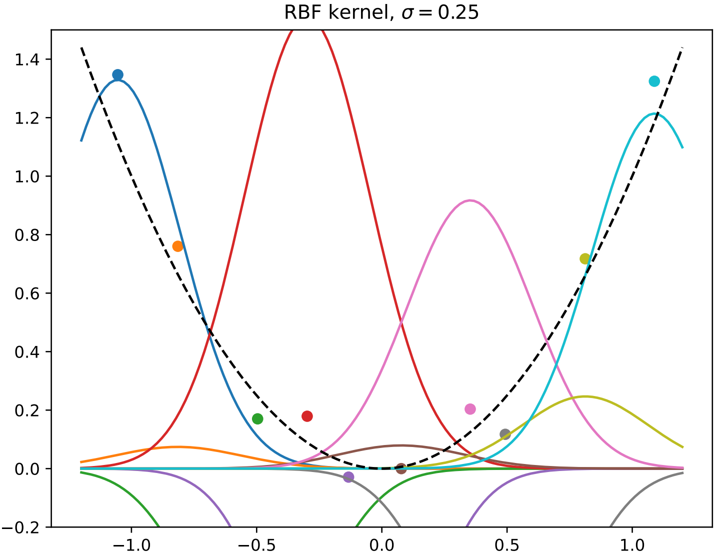
Example



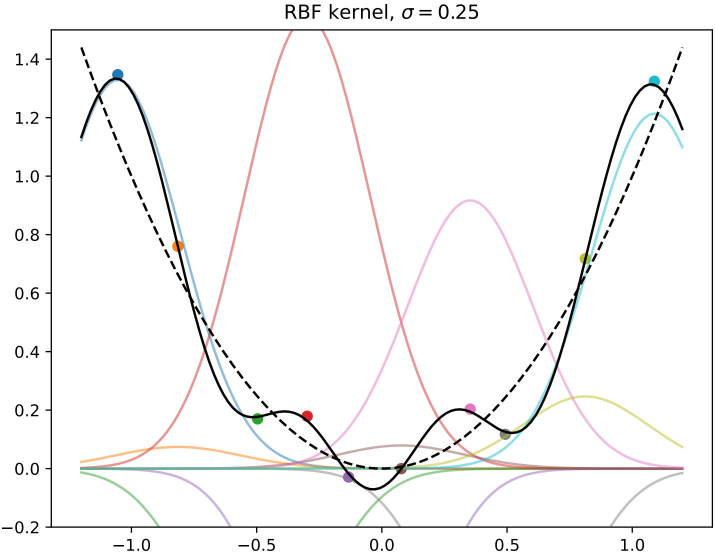
Example



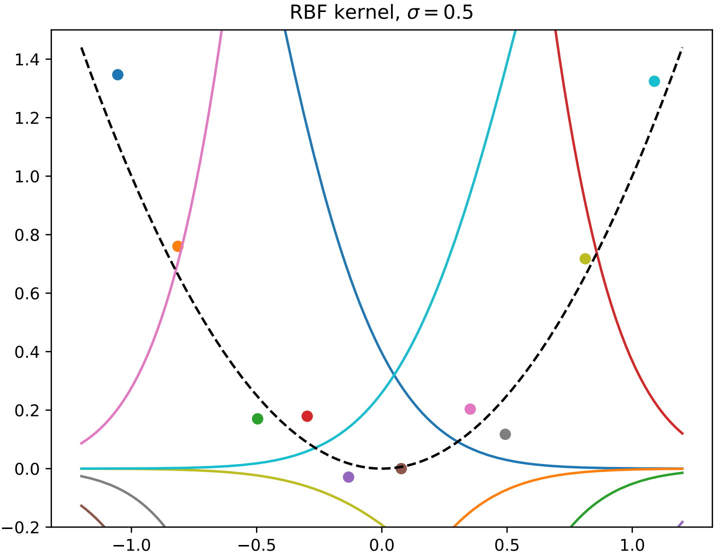
Example



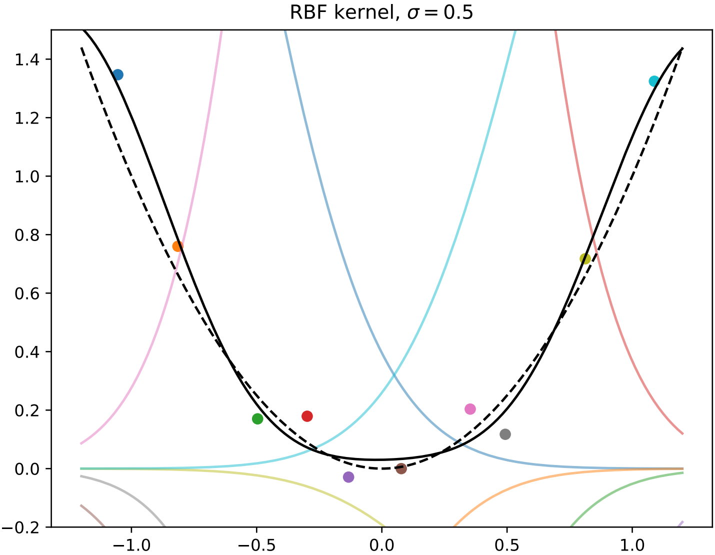
Example



Example



Example



Classification

Classification

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

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

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

Softmax

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

$$\begin{aligned} f(x) &= \mathrm{softmax}(g(x)); \\ f^i(x) &= \frac{\exp(g^i(x))}{\sum_k \exp(g^k(x))}. \end{aligned}$$

Logistic regression

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

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

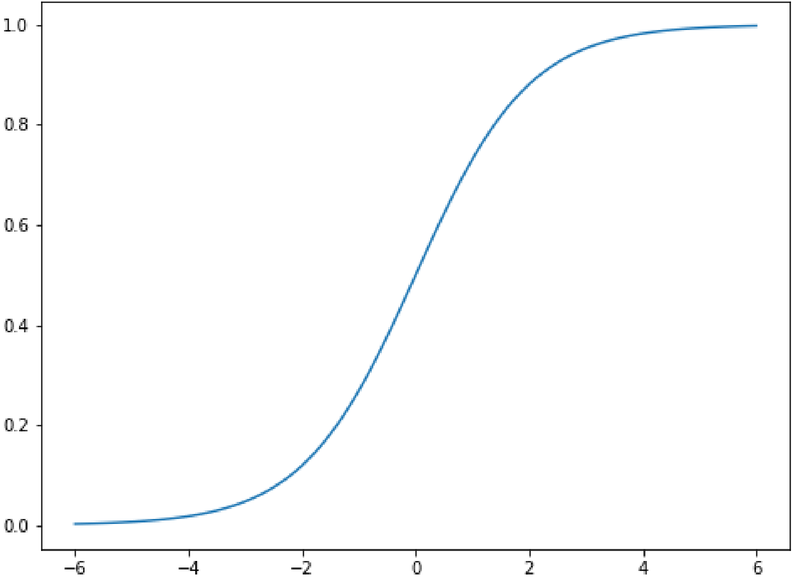
$$f_1(x) = \frac{\exp(w_1 \cdot x + b_1)}{\exp(w_1 \cdot x + b_1) + \exp(w_2 \cdot x + b_2)} =$$

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

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

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

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

$$\begin{aligned} f(\theta) &\to \min; \\ \theta^* = \arg\min_\theta f(\theta). \end{aligned}$$

$$\begin{aligned} \theta^{t+1} &= \theta^t - \alpha \nabla f(\theta^t); \\ \theta^t &\to \theta^*, t \to \infty; \end{aligned}$$

Gradient Descent

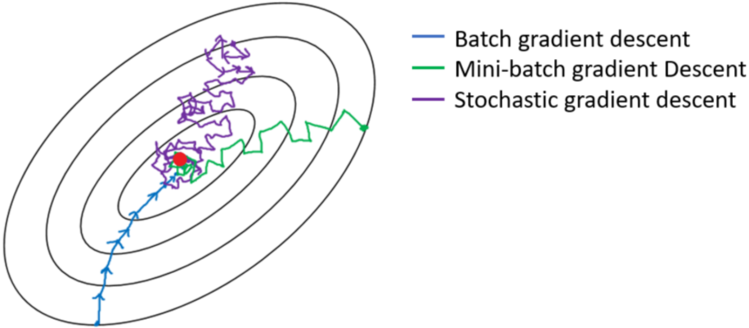
```
$\theta :=$ initialization
for $t := 1, \ldots$ do
    $\theta := \theta - \alpha \nabla f(\theta^t)$
end for
```

Stochastic Gradient Descent

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

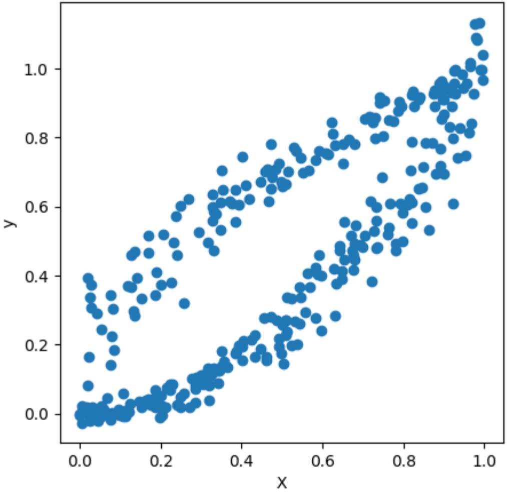
```
$\theta :=$ initialization
for $t := 1, \dots$ do
    $i := \mathrm{random}(1, \dots, N)$
    $\theta^{t+1} := \theta^t - \alpha \nabla f_i(\theta^t)$
end for
```

Illustration

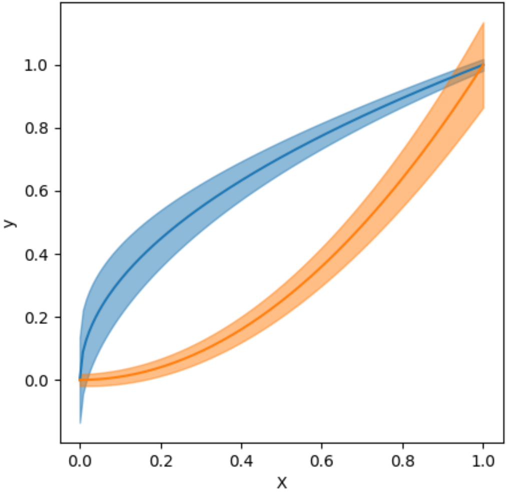


Source: `towardsdatascience.com`

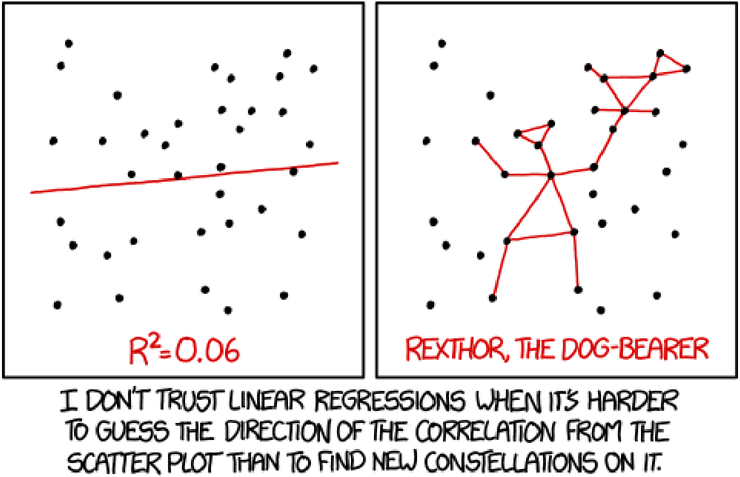
Tricky example



Tricky example



...


R²=0.06
REXTHOR, THE DOG-BEARER
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.

Source: `xkcd.com`