Artificial Intelligence and Machine Learning

Ivan Yamshchikov

Let’s talk business

Non-parametric models

— No functional form for $f(\cdot)$ is assumed
— The structure of the model is defined by the data
— May accurately fit a wider range of possible shapes for $f(\cdot)$

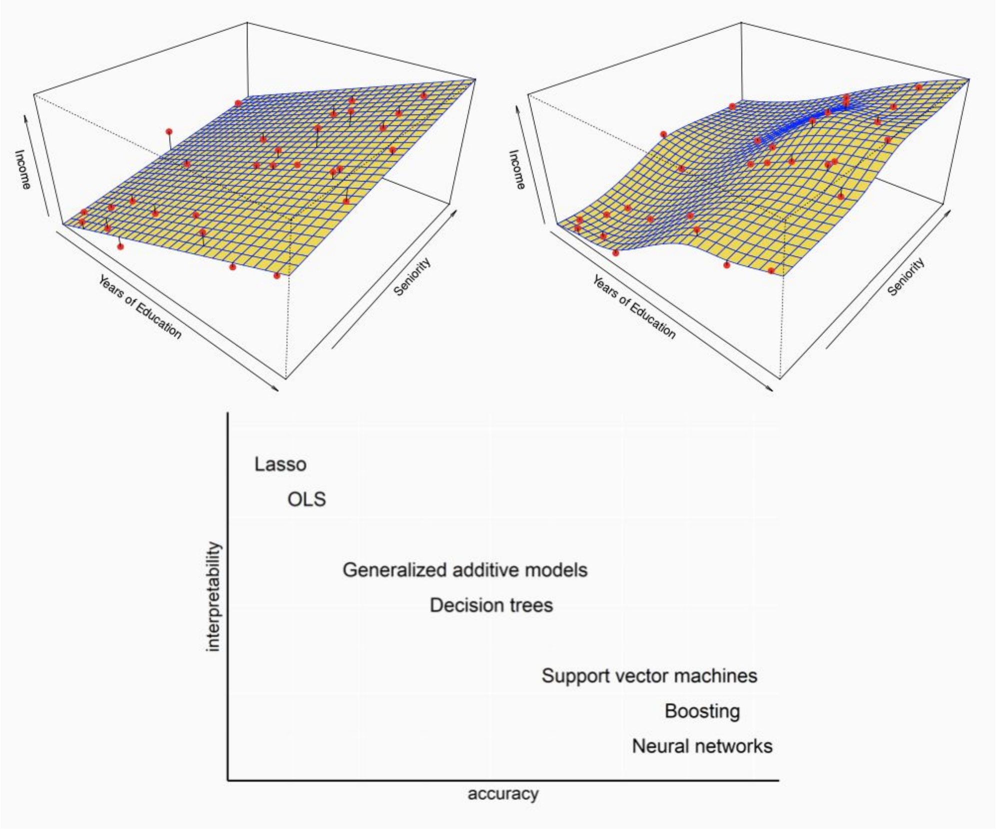
Disadvantage: a larger number of observations is required to obtain an accurate estimate of $f(\cdot)$; lower interpretability

Parametric models

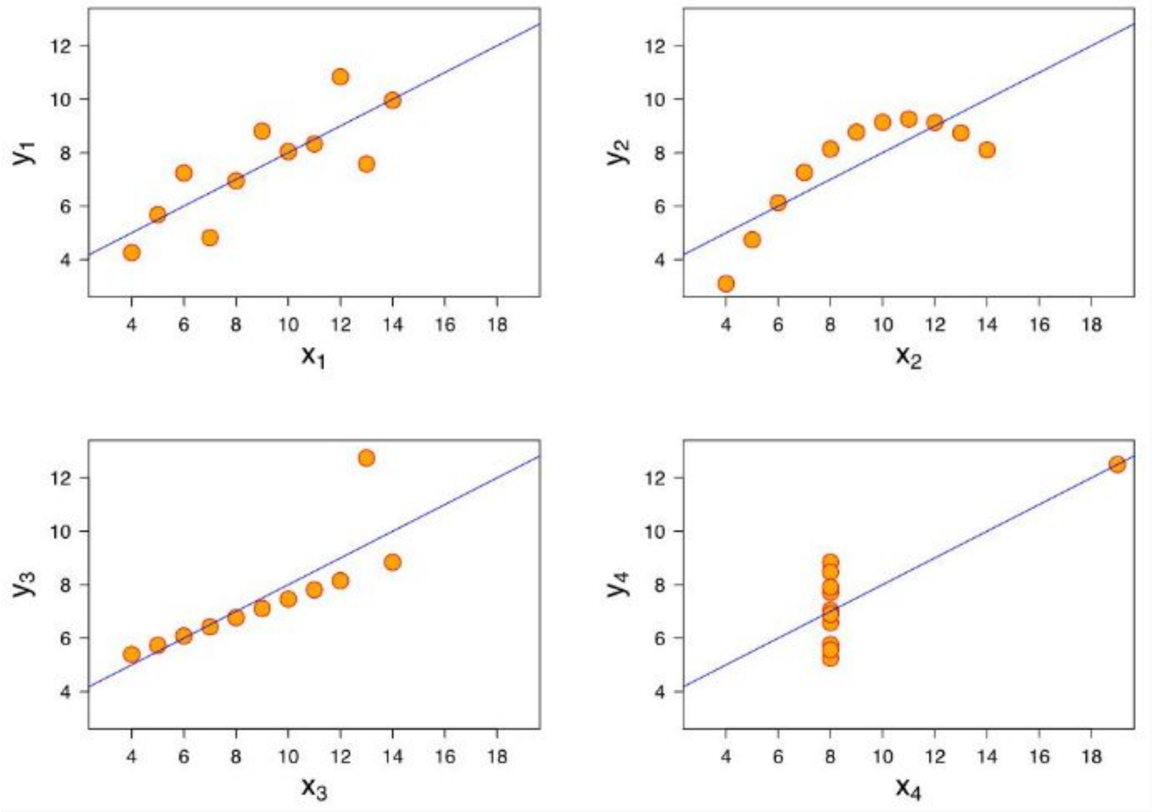
— We assume a functional form for f (·)
— Therefore, we reduce the problem of estimating f (·) down to one of estimating the model parameters/coefficients

Disadvantage: the functional form we choose may be very different from the true unknown f (·)

Trade-off between flexibility and interpretability



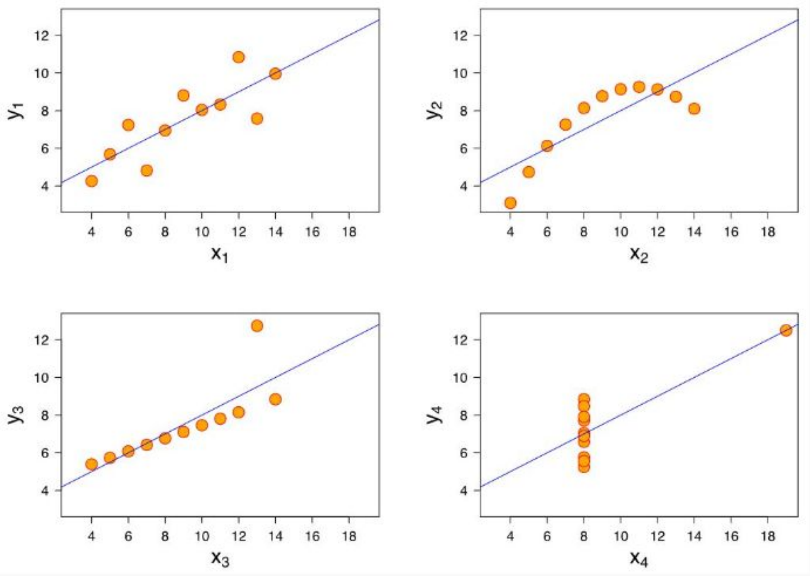
Are these data similar?



These datasets are similar “statistically”!

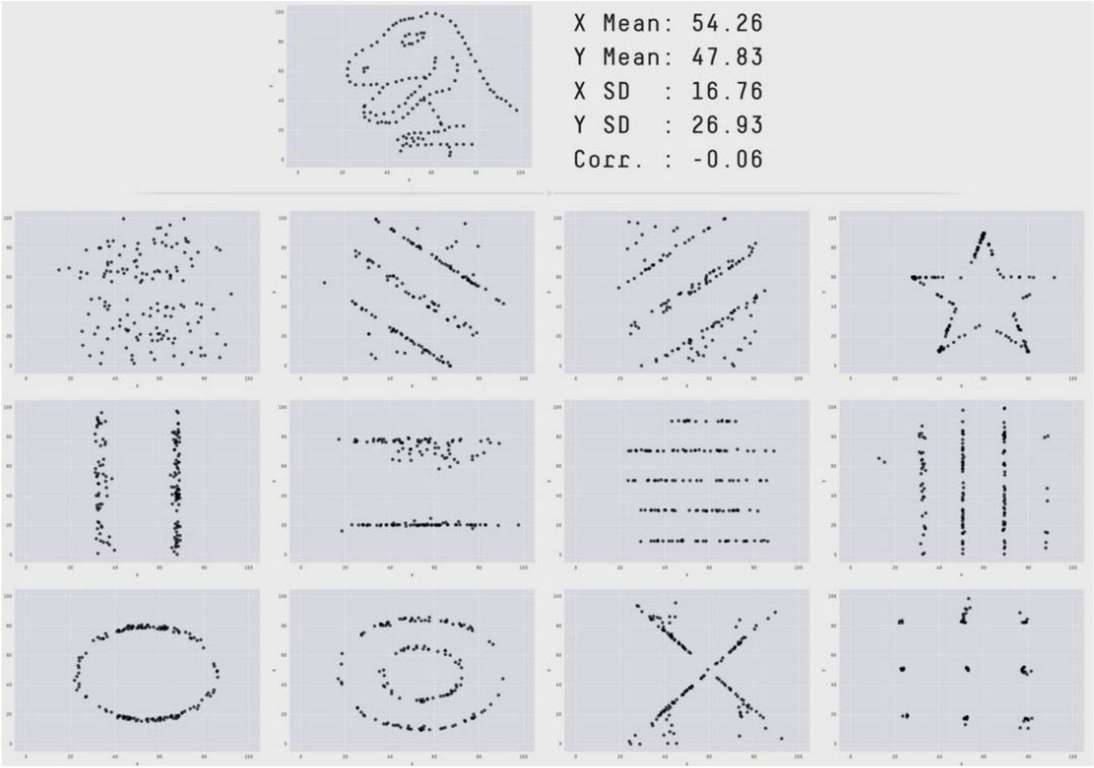
Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x	11	exact
Mean of y	7.5	to 2 decimal places
Sample variance of y	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression :	0.67	to 2 decimal places

Anscombe's quartet



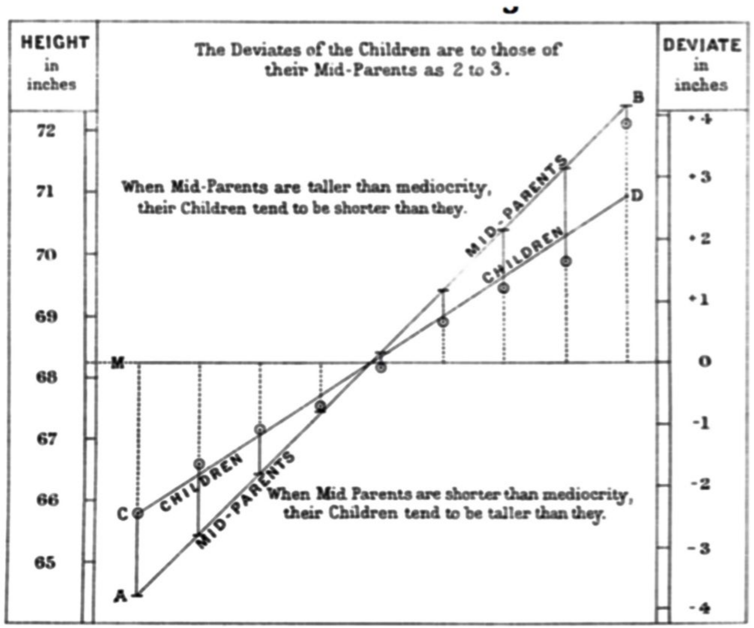
Property	Value	Accuracy
Mean of x	9	exact
Sample variance of x	11	exact
Mean of y	7.5	to 2 decimal places
Sample variance of y	4.125	±0.003
Correlation between x and y	0.816	to 3 decimal places
Linear regression line	$y = 3.00 + 0.500x$	to 2 and 3 decimal places, respectively
Coefficient of determination of the linear regression :	0.67	to 2 decimal places

Data — first!



Regression vs Classification

Linear regression



Linear regression

$$a(x) = w_0 + \sum_{j=1}^{d} w_j x^j,$$

$$a(x) = \sum_{j=1}^{d+1} w_j x^j = \langle w, x \rangle,$$

Can we solve this thing analytically?

Analytic solution

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

Is analytical solution a good one?

Linear regression

$$a(x) = w_0 + \sum_{j=1}^{d} w_j x^j,$$

$$a(x) = \sum_{j=1}^{d+1} w_j x^j = \langle w, x \rangle,$$

$$Q(w, x) = \frac{1}{\ell} \sum_{i=1}^{\ell} \left(\langle w, x_i \rangle - y_i \right)^2 .$$

What about quality?

$$Q(w, x) = \frac{1}{\ell} \sum_{i=1}^{\ell} \left(\langle w, x_i \rangle - y_i \right)^2 .$$

What about “learning”?

Learning

$$Q(w, x) = \frac{1}{\ell} \sum_{i=1}^{\ell} (\langle w, x_i \rangle - y_i)^2 \to \min_w .$$

$$Q(w, X) = \frac{1}{\ell} \|Xw - y\|^2 \to \min_w .$$

Gradient descent

$$w^t = w^{t-1} - \eta_t \nabla Q(w^{t-1}, X).$$

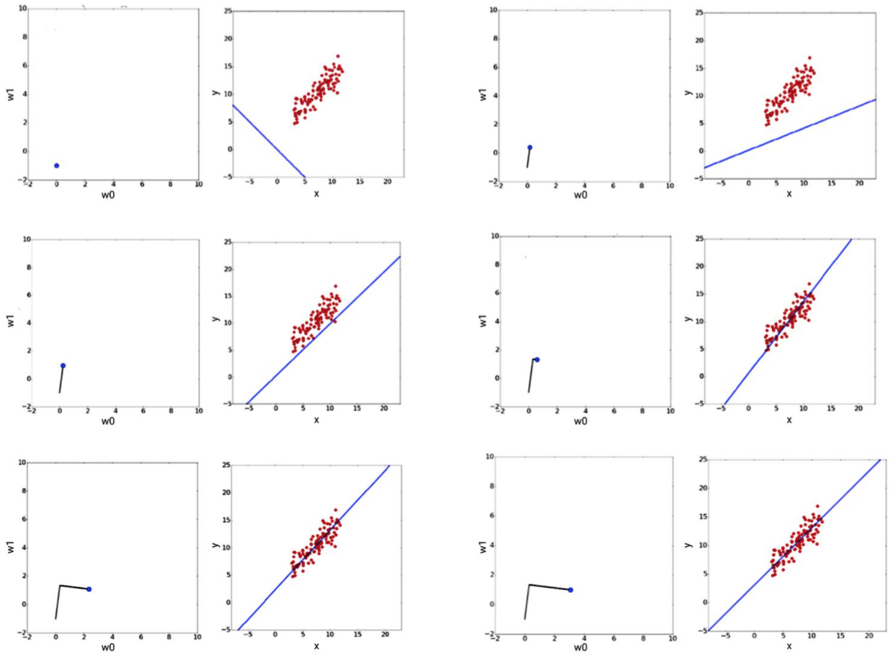
$$\|w^t - w^{t-1}\| < \varepsilon.$$

GD for Linear Regression

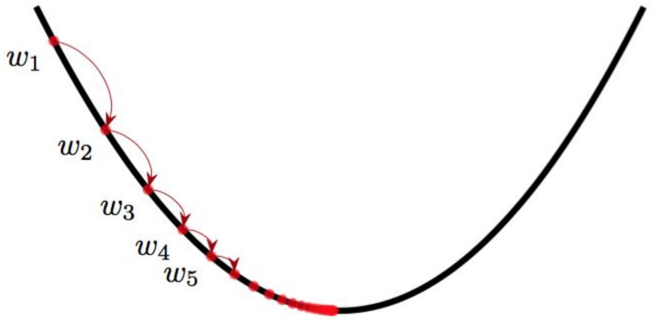
$$Q(w_0, w_1, X) = \frac{1}{\ell}\sum_{i=1}^{\ell}(w_1 x_i + w_0 - y_i)^2.$$

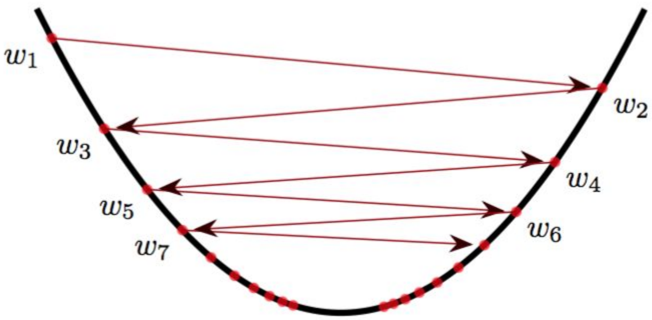
$$\frac{\partial Q}{\partial w_1} = \frac{2}{\ell}\sum_{i=1}^{\ell}(w_1 x_i + w_0 - y_i)x_i, \qquad \frac{\partial Q}{\partial w_0} = \frac{2}{\ell}\sum_{i=1}^{\ell}(w_1 x_i + w_0 - y_i).$$

Example



Learning rate matters





Multidimensional case is similar

$$Q(w, X) = \frac{1}{\ell}\|Xw - y\|^2 \to \min_{w},$$

$$\nabla_w Q(w, X) = \frac{2}{\ell}X^T(Xw - y)$$

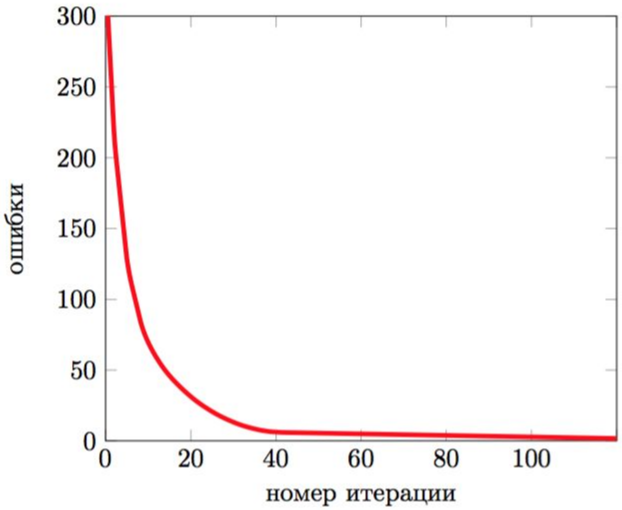
Is GD easy to compute?

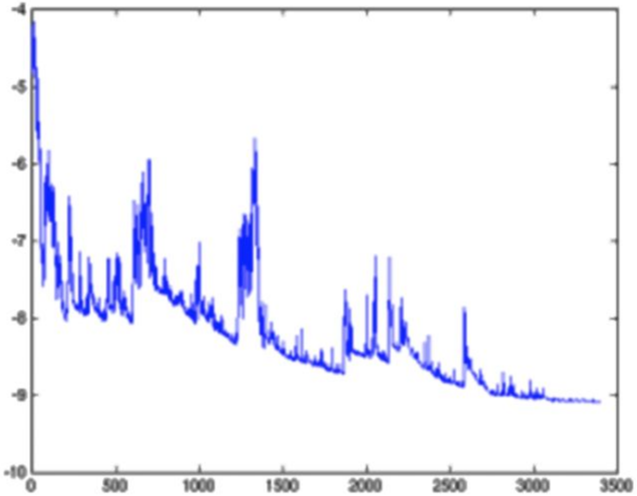
$$\frac{\partial Q}{\partial w_j} = \frac{2}{\ell} \sum_{i=1}^{\ell} x_i^j (\langle w, x_i \rangle - y_i).$$

Stochastic gradient descent

$$w^t = w^{t-1} - \eta_t \nabla Q(w^{t-1}, \{x_i\}).$$

Convergence of GD and SGD



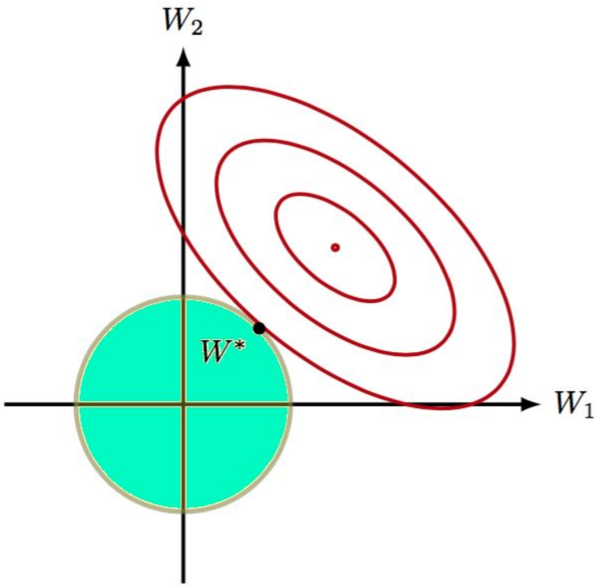


Regularization

$$\|w\|^2 = \sum_{j=1}^{d} w_j^2.$$

$$Q(w, X) + \lambda\|w\|^2 \to \min_w .$$

Regularization



Metrics

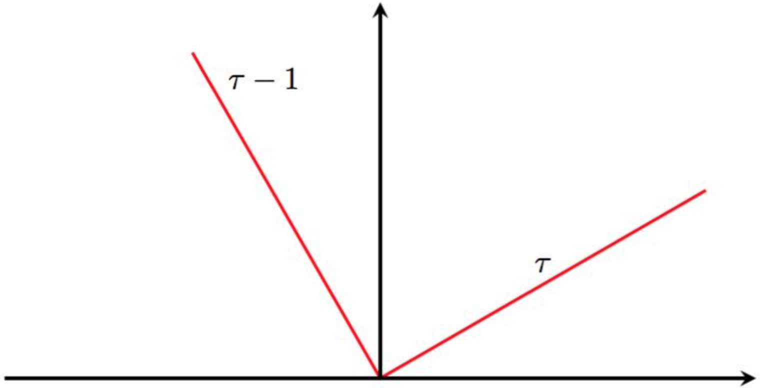
$$MSE(a, X) = \frac{1}{\ell}\sum_{i=1}^{\ell}(a(x_i) - y_i)^2\,.$$

$$MAE(a, X) = \frac{1}{\ell}\sum_{i=1}^{\ell}|a(x_i) - y_i|\,.$$

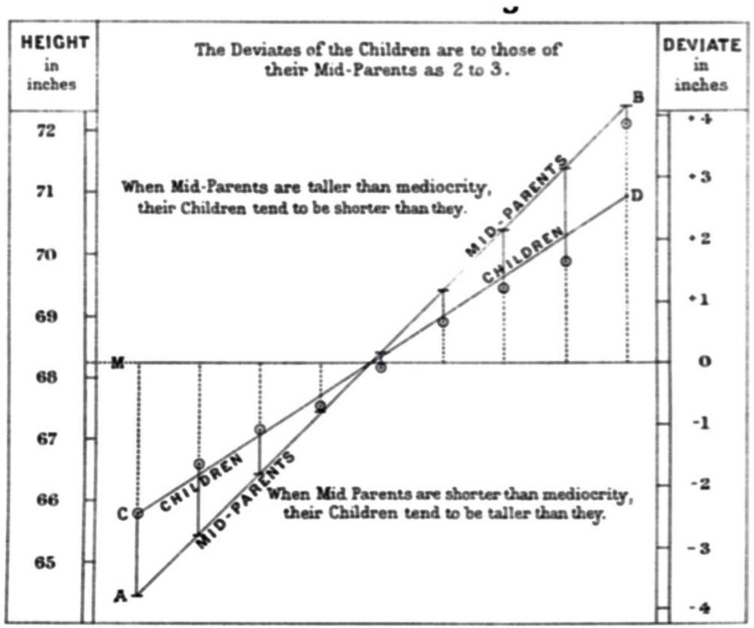
$$R^2(a, X) = 1 - \frac{\sum_{i=1}^{\ell}\left(a(x_i) - y_i\right)^2}{\sum_{i=1}^{\ell}(y_i - \bar{y})}, \qquad \bar{y} = \frac{1}{\ell}\sum_{i=1}^{\ell} y_i,$$

Quantile error

$$\rho_\tau(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} \Big((\tau - 1)\,[y_i < a(x_i)] + \tau\,[y_i \geq a(x_i)] \Big)\,(y_i - a(x_i))\,.$$



Linear regression statistically



Maximum likelihood estimator

$$X \sim F(x, \theta), \quad X^n = (X_1, ..., X_n),$$

$$L(X^n, \lambda) = \prod_{i=1}^{n} P(X = X_i, \theta).$$

$$\mathrm{argmax}_\lambda \ln L(X^N, \lambda)$$

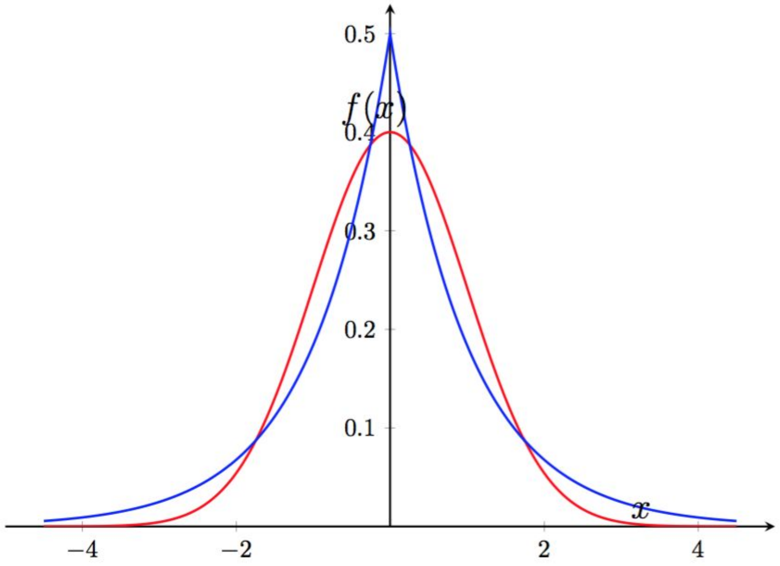
Gaussian noise

$$y = a(x) + \varepsilon,$$

$$a_* = \operatorname{argmin}_a \frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2$$

Laplacian noise

$$a_* = \text{argmin}_a \frac{1}{\ell} \sum_{i=1}^{\ell} |a(x_i) - y_i|. \qquad f(x) = \frac{\alpha}{2} e^{-\alpha|x|}.$$

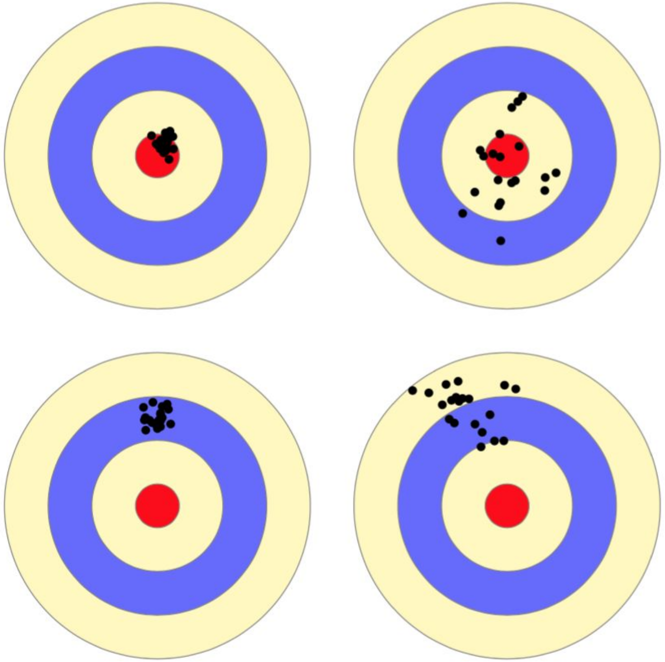


Ridge and Lasso regressions

$$w_* = \operatorname{argmin}_w \left(\frac{1}{\ell} \sum_{i=1}^{\ell} (\langle w, x_i \rangle - y_i)^2 + \lambda \sum_{j=1}^{d} w_j^2 \right).$$

$$w_* = \operatorname{argmin}_w \left(\frac{1}{\ell} \sum_{i=1}^{\ell} (\langle w, x_i \rangle - y_i)^2 + \lambda \sum_{j=1}^{d} |w_j| \right).$$

Dispersion and shift



Generalization of linear models

$$g\left(\mathbb{E}(y|x)\right) \approx \langle w, x \rangle,$$

$$\mathbb{E}(y|x) \approx g^{-1}\left(\langle w, x \rangle\right)$$