

# A Maximum-Likelihood View of Linear Regression

## Machine Learning (CSCI 5770G)

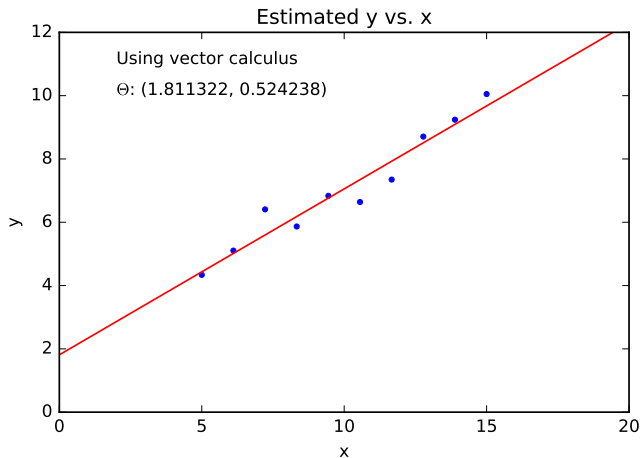
Faisal Z. Qureshi

<http://vclab.science.utoronto.ca>



# Probabilistic view of linear regression

We now turn our attention to probabilistic view of linear regression.



# Univariate Gaussian distribution

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right)$$

$\mu$  is the center of mass or *mean*

$\sigma^2$  is the variance

$\mu$  and  $\sigma^2$  are sufficient statistics

## Sampling from a Gaussian

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

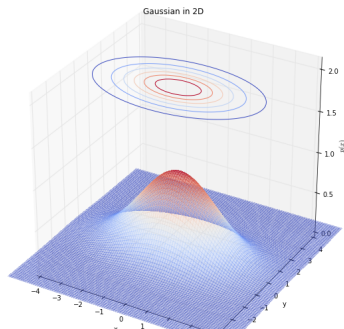
# Multivariate Gaussian distribution

Gaussian distribution in  $d$ -dimensions

$$\mathcal{N}(\mathbf{x}|\mu, \Sigma) = \frac{1}{|\Sigma|^{\frac{1}{2}} (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1}(\mathbf{x} - \mu)\right)$$

$\mathbf{x}, \mu \in \mathbb{R}^d$  and  $\Sigma \in \mathbb{R}^{d \times d}$

**Ex: 2D Gaussian**



## Covariance

Covariance between two random variables  $X$  and  $Y$  measures the degree to which these variables are linearly related.

$$\text{cov}[X, Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

$\mathbb{E}[X]$  is the *expected* value of the random variable  $X$ .

$$\mathbb{E}[X] = \int xp(x)dx = \mu$$

## Covariance matrix $\Sigma$

If  $\mathbf{x} \in \mathbb{R}^d$  random vector, its covariance matrix  $\Sigma$  is defined as follows:

$$\Sigma = \text{cov}[\mathbf{x}] = \begin{bmatrix} \text{var}[X_1] & \text{cov}[X_1, X_2] & \cdots & \text{cov}[X_1, X_d] \\ \text{cov}[X_2, X_1] & \text{var}[X_2] & \cdots & \text{cov}[X_2, X_d] \\ \vdots & \vdots & & \vdots \\ \text{cov}[X_d, X_1] & \text{cov}[X_d, X_2] & \cdots & \text{var}[X_d] \end{bmatrix}$$

## Likelihood example

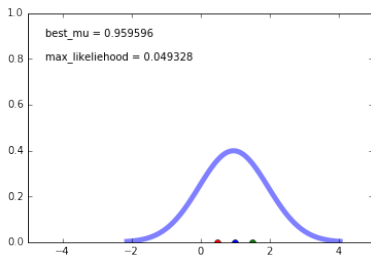
Consider the points:  $y_1 = 1$ ,  $y_2 = 0.5$  and  $y_3 = 1.5$ . The points are drawn from a Gaussian with unknown *mean*  $\theta$  and  $\sigma^2 = 1$ .

$$y_i \sim \mathcal{N}(\theta, 1).$$

Points are independent so

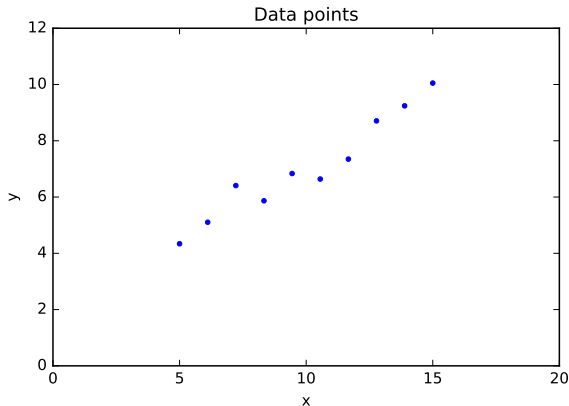
$$P(y_1, y_2, y_3 | \theta) = P(y_1 | \theta)P(y_2 | \theta)P(y_3 | \theta)$$

Our goal is to find the Gaussian (i.e., find its mean, since variance is already given) that maximizes the *likelihood* of this data.



## Linear regression

Consider data points  $(x^{(1)}, y^{(1)})$ ,  $(x^{(2)}, y^{(2)})$ ,  $\dots$ ,  $(x^{(N)}, y^{(N)})$ . Our goal is to learn a function  $f(x)$  that returns (predict) the value  $y$  given an  $x$ .





## Probabilistic view of linear regression

Let's assume that targets  $y^{(i)}$  are corrupted by Gaussian noise with 0 mean and  $\sigma^2$  variance

$$\begin{aligned}y^{(i)} &= (\theta_0 + \theta_1 x^{(i)}) + \mathcal{N}(0, \sigma^2) \\ &= \mathcal{N}(\theta_0 + \theta_1 x^{(i)}, \sigma^2)\end{aligned}$$

### Why assume Gaussian noise?

- ▶ Mathematically convenient
- ▶ A reasonably accurate assumption in practice
- ▶ Central Limit Theorem

### In higher dimensions

$$y^{(i)} = \mathcal{N}(\mathbf{x}^{(i)T} \boldsymbol{\theta}, \sigma^2)$$

## The likelihood for linear regression

Under the assumption that each  $y^{(i)}$  is independent and identically distributed (i.i.d.), we can write the likelihood of  $\mathbf{y}$  given data  $\mathbf{X}$  as follows:

$$\begin{aligned} p(\mathbf{y}|\mathbf{X}; \theta, \sigma) &= \prod_{i=1}^N p(y^{(i)}|\mathbf{x}^{(i)}; \theta, \sigma) \\ &= \prod_{i=1}^n \left(2\pi\sigma^2\right)^{-1/2} e^{-\frac{1}{2\sigma^2} \left(y^{(i)} - \mathbf{x}^{(i)T} \theta\right)^2} \\ &= \left(2\pi\sigma^2\right)^{-n/2} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n \left(y^{(i)} - \mathbf{x}^{(i)T} \theta\right)^2} \\ &= \left(2\pi\sigma^2\right)^{-n/2} e^{-\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\theta)^T (\mathbf{y} - \mathbf{X}\theta)} \end{aligned}$$

Aside: the “;” above indicate that we are following the *frequentist* approach, and we do not treat  $\theta$  as a random variable. Rather we view  $\theta$  as having some true value that we are trying to estimate.

# Probabilistic view of linear regression

Least squares loss

$$C(\theta) = (\mathbf{y} - \mathbf{X}\theta)^T (\mathbf{y} - \mathbf{X}\theta)$$

Probabilistic view

$$p(\mathbf{y}|\mathbf{X}; \theta, \sigma) = \left(2\pi\sigma^2\right)^{-n/2} e^{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\theta)^T(\mathbf{y}-\mathbf{X}\theta)}$$

Probability of data given parameters is related to the loss for linear regression that we obtained before.

# Maximum Likelihood Estimation (MLE)

## Likelihood

$$p(\mathbf{y}|\mathbf{X}; \theta, \sigma) = \left(2\pi\sigma^2\right)^{-n/2} e^{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\theta)^T(\mathbf{y}-\mathbf{X}\theta)}$$

## Negative log-likelihood

$$\begin{aligned} -\log p(\mathbf{y}|\mathbf{X}; \theta, \sigma) &= -\log \left( \left(2\pi\sigma^2\right)^{-n/2} e^{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\theta)^T(\mathbf{y}-\mathbf{X}\theta)} \right) \\ &= \frac{n}{2} \log \left(2\pi\sigma^2\right) + \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\theta)^T(\mathbf{y} - \mathbf{X}\theta) \end{aligned}$$

## Maximum likelihood estimation

The maximum likelihood estimate (MLE) of  $\theta$  is obtained by maximizing  $p(\mathbf{y}|\mathbf{X}; \theta, \sigma)$

$$\begin{aligned}\theta_{\text{ML}} &= \arg \max_{\theta} p(\mathbf{y}|\mathbf{X}; \theta, \sigma) \\ &= \arg \min_{\theta} -\log p(\mathbf{y}|\mathbf{X}; \theta, \sigma) \\ &= \arg \min_{\theta} \frac{n}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\theta)^T (\mathbf{y} - \mathbf{X}\theta) \\ &= \arg \min_{\theta} (\mathbf{y} - \mathbf{X}\theta)^T (\mathbf{y} - \mathbf{X}\theta)\end{aligned}$$

## Making predictions using MLE

For a previously unseen data  $\mathbf{x}^*$ , the target  $y^*$  can be obtained as follows:

$$y^* \sim \mathcal{N}(\theta_{\text{ML}}^T \mathbf{x}^*, \sigma^2)$$

## Kullback-Leibler (KL) divergence

Kullback-Leibler divergence is a measure of how much two probability distributions diverge from each other.

$$\begin{aligned}D_{\text{KL}}(p\|q) &= \int p(x) \log \frac{p(x)}{q(x)} dx \\ &= \mathbb{E}_{x \sim p(x)} \left[ \log \frac{p(x)}{q(x)} \right]\end{aligned}$$

For discrete probability distributions

$$D_{\text{KL}}(P\|Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

KL divergence is **not a measure of distance**, since it is not symmetric

$$D_{\text{KL}}(P\|Q) \neq D_{\text{KL}}(Q\|P)$$

## MLE and KL divergence

Consider the setting where we are attempting to fit a distribution  $P(x|\theta)$  to data that is drawn from some true distribution  $P(x|\theta^*)$ . One way to do so is to find the parameter  $\theta$  that minimizes the KL divergence between the two distributions.

$$\begin{aligned}\theta_{\min\text{KL}} &= \arg \min_{\theta} D_{\text{KL}} [P(x|\theta^*) || P(x|\theta)] \\ &= \arg \min_{\theta} \mathbb{E}_{x \sim P(x|\theta^*)} \left[ \log \frac{P(x|\theta^*)}{P(x|\theta)} \right] \\ &= \arg \min_{\theta} \mathbb{E}_{x \sim P(x|\theta^*)} \left[ \underbrace{\log P(x|\theta^*)}_{\text{does not effect minima}} - \log P(x|\theta) \right] \\ &= \arg \min_{\theta} \mathbb{E}_{x \sim P(x|\theta^*)} [-\log P(x|\theta)] \\ &= \arg \max_{\theta} \mathbb{E}_{x \sim P(x|\theta^*)} [\log P(x|\theta)] \\ &\quad \underbrace{\hspace{10em}}_{\text{MLE}}\end{aligned}$$



## MLE and KL divergence

It turns out that for i.i.d. (independent, identically distributed) data from a some (unknown true) distribution MLE minimizes the KL divergence.

## Ridge regression and Bayes rule

Previously we saw the loss function for ridge regression

$$C(\theta) = (\mathbf{y} - \mathbf{X}\theta)^T (\mathbf{y} - \mathbf{X}\theta) + \delta^2 \theta^T \theta$$

We can cast the above in probabilistic terms

$$p(y|\mathbf{x}, \theta) = \frac{1}{Z_1} e^{-((\mathbf{y} - \mathbf{X}\theta)^T (\mathbf{y} - \mathbf{X}\theta))}$$

Then

$$p(\theta) = \frac{1}{Z_2} e^{-\delta^2 \theta^T \theta}$$

becomes *prior*.

# Summary

- ▶ We developed a probabilistic view of linear regression.
- ▶ Maximum likelihood estimation
- ▶ Kullback-Leibler divergence
- ▶ Relationship between MLE and KL

# Copyright and License

©Faisal Z. Qureshi



This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.