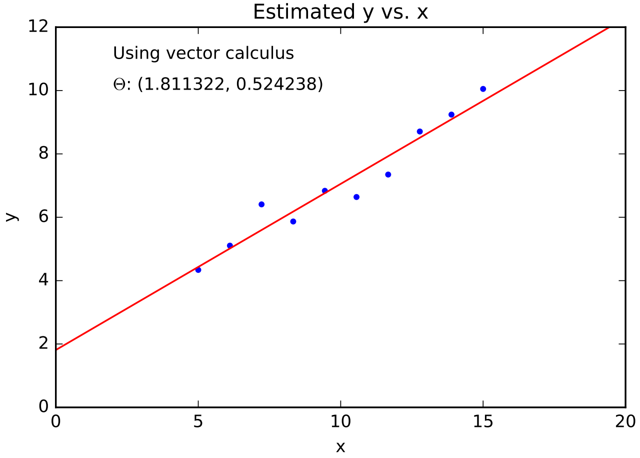
Maximum Likelihood

Topics in Digital Media

Faisal Qureshi

Probabilistic view of linear regression

We now turn our attention to probabilistic view of linear regression



Univariate Gaussian distribution

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

μ is the center of mass or *mean*

σ^2 is the variance

μ and σ^2 are sufficient statistics

Sampling from a Gaussian

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

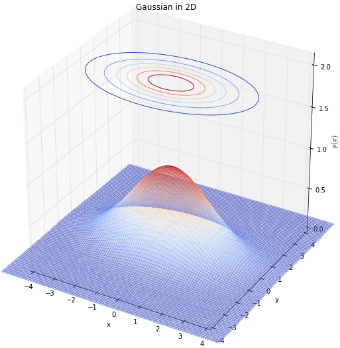
Multivariate Gaussian distribution

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 $\Sigma \in \mathbb{R}^{d \times d}$

Example: Gaussian in 2D



Covariance

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

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

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

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

Covariance matrix Σ

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

$$\Sigma = \mathrm{cov}[\mathbf{x}] = \left[\begin{array}{cccc} \mathrm{var}[X_1] & \mathrm{cov}[X_1, X_2] & \cdots & \mathrm{cov}[X_1, X_d] \\ \mathrm{cov}[X_2, X_1] & \mathrm{var}[X_2] & \cdots & \mathrm{cov}[X_2, X_d] \\ \vdots & \vdots & & \vdots \\ \mathrm{cov}[X_d, X_1] & \mathrm{cov}[X_d, X_2] & \cdots & \mathrm{var}[X_d] \end{array} \right]$$

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* θ and $\sigma^2 = 1$.

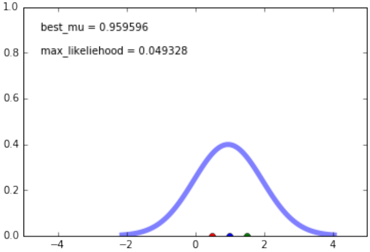
$$y_i \sim \mathcal{N}(\theta, 1)$$

.

Points are independent so

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

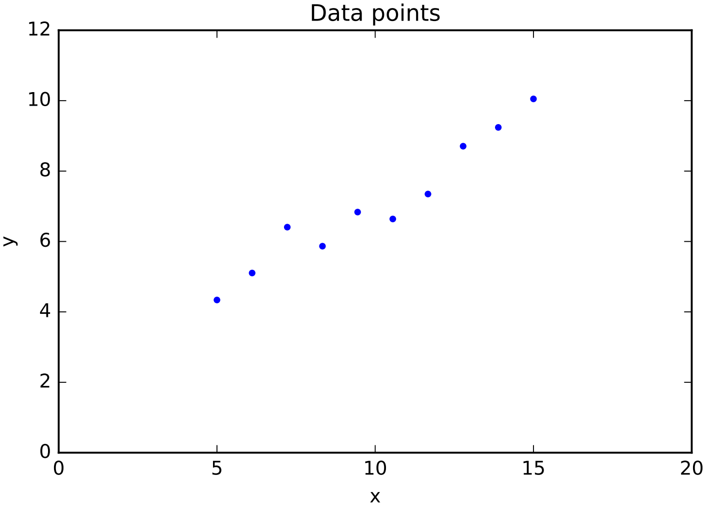
Our goal is to find the Gaussian (i.e., find its mean, since variance is already given) that maximizes the *likelihood* of this data.



From Nando de Freitas

Linear regression

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



The likelihood for linear regression

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

$$\begin{aligned} y^{(i)} &= \theta^T x^{(i)} + \mathcal{N}(0, \sigma^2) \\ &= \mathcal{N}\left(\theta^T x^{(i)}, \sigma^2\right) \end{aligned}$$

In higher dimensions, we write:

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

Why assume Gaussian noise?

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

The likelihood for linear regression

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

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

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

Probability of data given parameters

Loss for linear regression

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

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

Maximum likelihood estimation (1)

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

$$\theta_{\mathrm{ML}} = \arg\max_{\theta} \prod_{i=1}^{N} p(y^{(i)}|\mathbf{x}^{(i)}; \theta, \sigma)$$

Log likelihood

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

Maximum likelihood estimation (2)

Making predictions using MLE

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

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

Entropy

Entropy H is a measure of uncertainty associated with a random variable.

$$H(X) = -\sum_x p(x|\theta) \log p(x|\theta)$$

Example

Entropy of a Gaussian in D dimensions

$$H\left(\mathcal{N}(\mu, \Sigma)\right) = \frac{1}{2} \ln \left[(2\pi e)^D |\Sigma| \right]$$

Kullback-Leibler divergence

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

For discrete probability distributions

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

For continuous probability distributions

$$D_{KL}\left(P \| Q\right) = \int p(x) \log \frac{p(x)}{q(x)} dx$$

Kullback-Leibler divergence

MLE : For IID data from some distribution $p(x|\theta_0)$ the MLE minimizes the KL divergence (KULLBACK - LEIBLER Divergence)

$$\hat{\theta} = \arg\max_{\theta} \prod_{i=1}^{m} p(x^{(i)}|\theta)$$

$$= \arg\max_{\theta} \sum_{i=1}^{m} \log p(x^{(i)}|\theta)$$

$$= \arg\max_{\theta} \frac{1}{m}\sum_{i=1}^{m} \log p(x^{(i)}|\theta) - \frac{1}{m}\sum_{i=1}^{m} \log p(x^{(i)}|\theta_0)$$

$$= \arg\max_{\theta} \frac{1}{m}\sum_{i=1}^{m} \log p(x^{(i)}|\theta) \Big/ p(x^{(i)};\theta_0)$$

$$= \arg\min_{\theta} \frac{1}{m}\sum_{i=1}^{m} \log p(x^{(i)}|\theta_0) / p(x^{(i)}|\theta)$$

$$\approx \arg\min \int \log \frac{p(x|\theta_0)}{p(x|\theta)} \, p(x|\theta_0) \, dx$$

MLE and KL divergence

It turns out that for i.i.d. (independant, identically distributed) data from a some (unknown true) distribution $p(x|\theta_{\mathrm{True}})$ MLE minimizes the *Kullback-Leibler* (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

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

Then

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

becomes *prior*.

Summary

- We developed a probabilistic view of linear regression.