

Regression

Regression is prediction of continuous outputs.
The statistical approach to regression is to learn and apply a probability function.

$P(x, y)$ = joint probability function

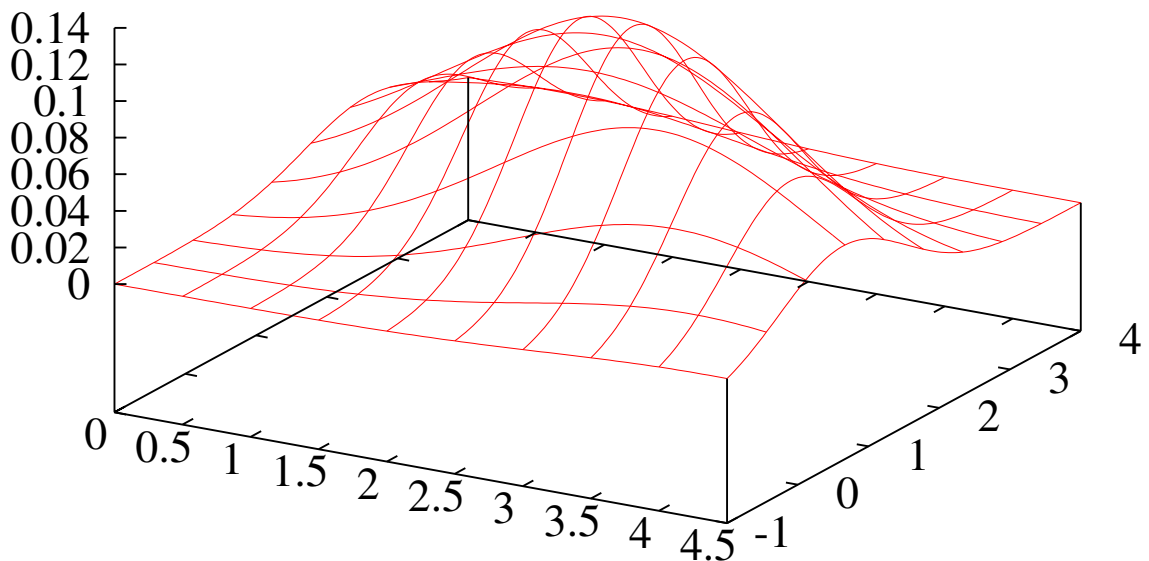
$P(y | x)$ = conditional probability function

Example: A bivariate normal distribution with mean $\boldsymbol{\mu} = (2.5, 1.5)$ and covariance matrix:

$$\boldsymbol{\Sigma} = \begin{bmatrix} 2.1 & -0.9 \\ -0.9 & 1.2 \end{bmatrix}$$

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

$P(x, y)$ ———



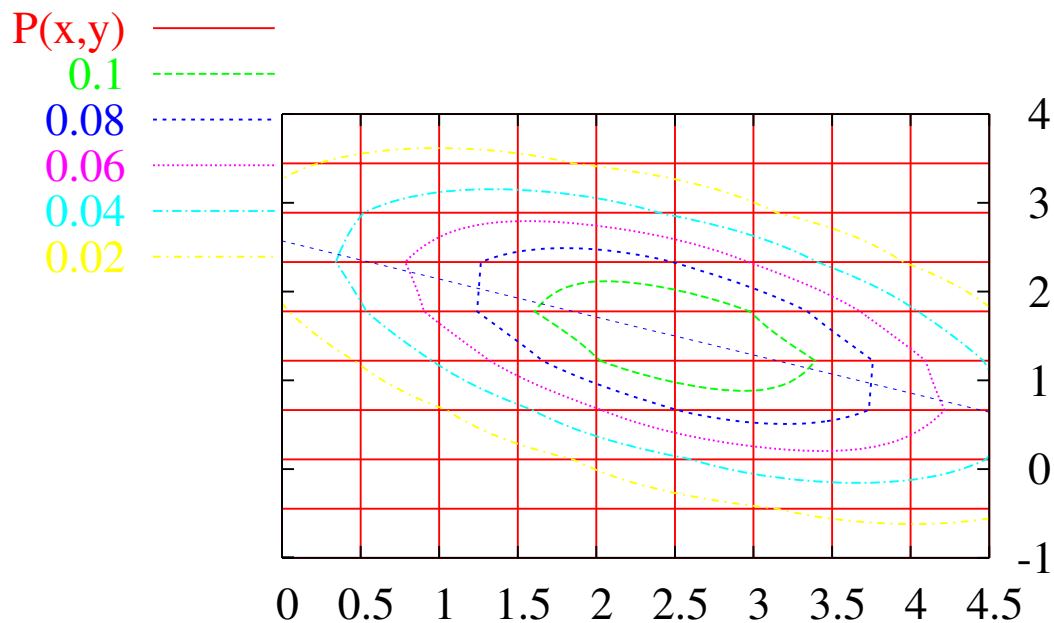
Prediction Given $P(\mathbf{x}, y)$

When the probability distribution $P(\mathbf{x}, y)$ is known, squared error is minimized by predicting $\mathbb{E}(y \mid \mathbf{x})$, the expected value of the output y given the inputs \mathbf{x} .

$$P(y \mid \mathbf{x}) = \frac{P(\mathbf{x}, y)}{P(\mathbf{x})} = \frac{P(\mathbf{x}, y)}{\int_{-\infty}^{+\infty} P(x, y) dy}$$

$$\mathbb{E}(y \mid \mathbf{x}) = \int_{-\infty}^{+\infty} y P(y \mid x) dy$$

This shows the prediction for the example normal distribution.



Learning Normal Distributions

For a normal distribution with mean (μ_x, μ_y) and covariance matrix:

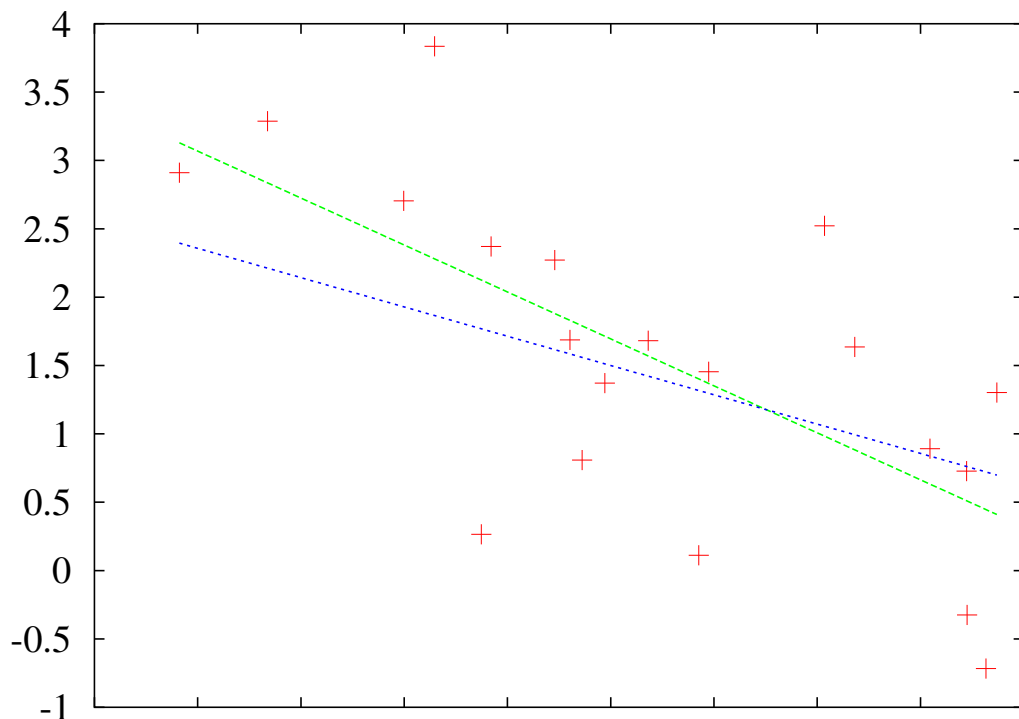
$$\begin{bmatrix} \sigma_x^2 & \sigma_{xy}^2 \\ \sigma_{xy}^2 & \sigma_y^2 \end{bmatrix}$$

then y can be predicted by:

$$\mu_y + \frac{\sigma_{xy}^2}{\sigma_x^2}(x - \mu_x)$$

However, in learning, we are given a sample of points, not the true distribution. I used the example distribution to generate 20 points.

This results in learning $3.412 - 0.687x$, while the target is $2.571 - 0.429x$.



Classification

Classification is prediction of categories or classes. The statistical approach to classification assumes that each class ω_i has its own probability distribution $P(\mathbf{x} \mid \omega_i)$.

The Bayesian decision criterion is to determine:

$$\arg \max_i P(\omega_i \mid \mathbf{x})$$

the class with the maximum *a posteriori* (MAP) probability. More generally, we want the class that minimizes *loss* or *risk*.

Bayesian classification needs to determine $P(\omega_i \mid \mathbf{x})$ from $P(\mathbf{x} \mid \omega_i)$. By Bayes' theorem:

$$P(\omega_i \mid \mathbf{x}) = \frac{P(\mathbf{x}, \omega_i)}{P(\mathbf{x})} = \frac{P(\mathbf{x} \mid \omega_i) P(\omega_i)}{P(\mathbf{x})}$$

$P(\omega_i)$ is the prior probability of class ω_i , i.e., how often this class occurs.

We prefer ω_1 over ω_2 if $P(\omega_1 \mid \mathbf{x}) > P(\omega_2 \mid \mathbf{x})$. This is equivalent to determining if:

$$P(\mathbf{x} \mid \omega_1) P(\omega_1) > P(\mathbf{x} \mid \omega_2) P(\omega_2) \quad (1)$$

because $P(\mathbf{x})$ is a common, positive term.

Decision Making

Inequality (1) can be rewritten in different ways:

$$\frac{P(\mathbf{x} | \omega_1) P(\omega_1)}{P(\mathbf{x} | \omega_2) P(\omega_2)} > 1$$

$$\ln \frac{P(\mathbf{x} | \omega_1) P(\omega_1)}{P(\mathbf{x} | \omega_2) P(\omega_2)} > \ln 1$$

$$\ln \frac{P(\mathbf{x} | \omega_1)}{P(\mathbf{x} | \omega_2)} + \ln \frac{P(\omega_1)}{P(\omega_2)} > 0$$

The last condition illustrates log-likelihoods.

Let L_{ij} = loss of deciding ω_i when ω_j is true.
 L_{ij} is the extra cost or risk of being wrong.

Bayes' risk criterion is to choose:

$$\arg \min_i \sum_j L_{ij} P(\omega_j | \mathbf{x})$$

the class that minimizes loss.

Assuming two classes and $L_{ii} = 0$, choose ω_1 if:

$$L_{21} P(\omega_1 | \mathbf{x}) > L_{12} P(\omega_2 | \mathbf{x}) \quad (2)$$

This is equivalent to:

$$L_{21} P(\mathbf{x} | \omega_1) P(\omega_1) > L_{12} P(\mathbf{x} | \omega_2) P(\omega_2)$$

Inequality (2) is also equivalent to:

$$\ln \frac{L_{21}}{L_{12}} + \ln \frac{P(\mathbf{x} | \omega_1)}{P(\mathbf{x} | \omega_2)} + \ln \frac{P(\omega_1)}{P(\omega_2)} > 0$$

0-1 loss is defined as $L_{ii} = 0$ otherwise $L_{ij} = 1$.

Note that Bayes' risk criterion with 0-1 loss is equivalent to MAP.