EC-332 Machine Learning

Nazar Khan

Department of Computer Science
University of the Punjab

Maximum Likelihood Estimation

In this lecture . . .

- Gaussian distribution
- Gaussian density estimation
- Probabilistic polynomial curve fitting

Gaussian Distribution
Univariate

- Known as the queen of distributions.
- Also called the *Normal distribution* since it models the distribution of almost all natural phenomenon.
- For continuous variables.

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

  where μ is the *mean*, σ^2 is the *variance* and σ is the *standard deviation*.
- Reciprocal of variance, $\beta = \frac{1}{\sigma^2}$ is called *precision*.

Gaussian Distribution
Univariate

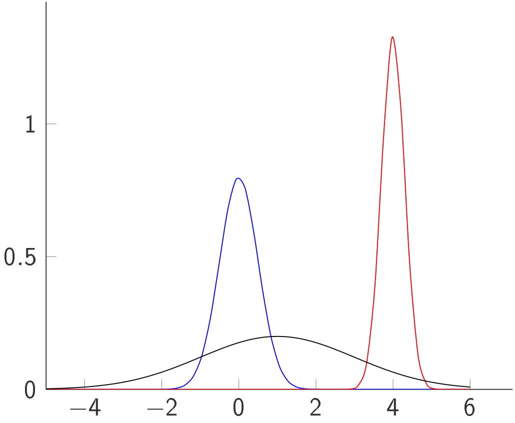

Figure: Plots of $\mathcal{N}(0, 0.5^2)$, $\mathcal{N}(4, 0.3^2)$ and $\mathcal{N}(1, 2^2)$. Notice that density is not the same as probability and can be greater than 1.

Gaussian Distribution
Multivariate

- Multivariate form for $D-$dimensional vector $\mathbf{x}$ of continuous variables

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

where the $D \times D$ matrix $\mathbf{\Sigma}$ is called the *covariance matrix* and $|\mathbf{\Sigma}|$ is its determinant.

Gaussian Distribution
Multivariate

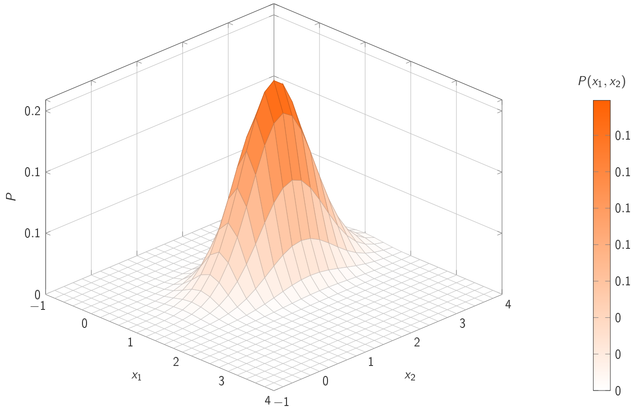

Figure: Plot of bivariate Gaussian distribution with mean $\mu = (1, 2)^T$ and $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{4} \end{bmatrix}$.

Gaussian Distribution
Multivariate

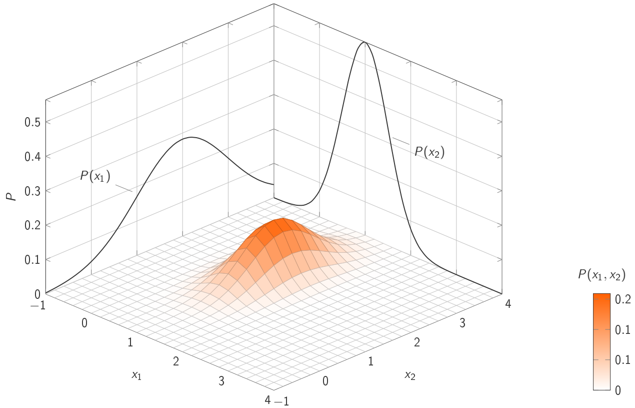

Figure: Plot of bivariate Gaussian distribution with mean $\boldsymbol{\mu} = (1, 2)^T$ and $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{4} \end{bmatrix}$. Marginal distributions $p(x_1)$ and $p(x_2)$ are also shown.

Independent and Identically Distributed

- Let $\mathcal{D} = (x_1, \ldots, x_N)$ be a set of N random numbers.
- If value of any x_i does not affect the value of any other x_j, then the x_is are said to be *independent*.
- If each x_i follows the same distribution, then the x_is are said to be *identically distributed*.
- Both properties combined are abbreviated as *i.i.d*.
- Assuming the x_is are i.i.d under $\mathcal{N}(\mu, \sigma^2)$

$$p(\mathcal{D}|\mu, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \sigma^2)$$

- This is known as the *likelihood function* for the Gaussian.
  - Likelihood of observed data given the Gaussian model with parameters (μ, σ^2).

The Log Function

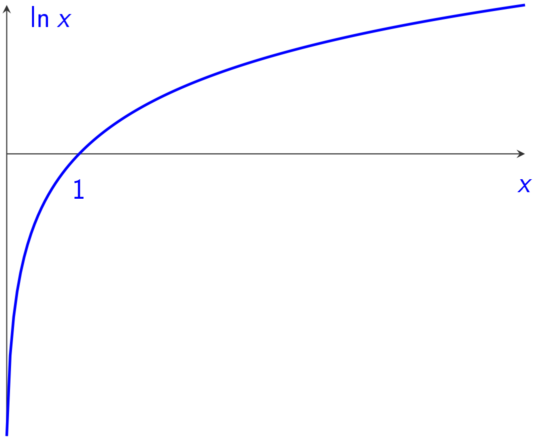

Figure: The log function is a monotonically increasing function. If $x_1 > x_2$, then $\log(x_1) > \log(x_2)$.

Fitting a Gaussian

- Assuming we have i.i.d data $\mathcal{D} = (x_1, \ldots, x_N)$, how can we find the parameters of the Gaussian distribution that generated it?
- Find the (μ, σ^2) that *maximise the likelihood*. This is known as the *maximum likelihood (ML)* approach.
- Since logarithm is a monotonically increasing function, maximising the log of a function is equivalent to maximising the function.
- Logarithm of the Gaussian
  - is a simpler function, and
  - is numerically superior (consider taking product of very small probabilities versus taking the sum of their logarithms).

Log Likelihood

- Log likelihood of Gaussian becomes

$$\ln p(\mathcal{D}|\mu, \sigma^2) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)$$

- Maximising w.r.t μ, we get

$$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

- Maximising w.r.t σ^2, we get

$$\sigma^2_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{ML})^2$$

Bias of Maximum Likelihood

- Since $\mathbb{E}[\mu_{ML}] = \mu$, ML estimates the mean correctly.
- But since $\mathbb{E}\left[\sigma^2_{ML}\right] = \left(\frac{N-1}{N}\right)\sigma^2$, ML underestimates the variance by a factor $\frac{N-1}{N}$.
- This phenomenon is called *bias* and lies at the root of over-fitting.

Polynomial Curve Fitting

A Probabilistic Perspective

- Our earlier treatment of curve fitting was via error minimization.
- Now we take a probabilistic perspective.
- The real goal: make accurate prediction t for new input x given training data $(\mathbf{x}, \mathbf{t})$.
- Prediction implies uncertainty. Therefore, target value can be modelled via a probability distribution.
- We assume that given x, the target variable t has a Gaussian distribution.

$$\begin{aligned} p(t|x, \mathbf{w}, \beta) &= \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1}) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(t - y(x, \mathbf{w}))^2\right\} \end{aligned} \tag{1}$$

Polynomial Curve Fitting

A Probabilistic Perspective

- Knowns: Training set $(\mathbf{x}, \mathbf{t})$.
- Unknowns: Parameters $\mathbf{w}$ and β.
- Assuming training data is i.i.d likelihood function becomes

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|y(x_n, \mathbf{w}), \beta^{-1})$$

- Log of likelihood becomes

$$\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = -\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 - \frac{N}{2}\ln \beta^{-1} - \frac{N}{2}\ln(2\pi)$$

- Maximization of likelihood w.r.t $\mathbf{w}$ is equivalent to minimization of $\frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2$.

Polynomial Curve Fitting
A Probabilistic Perspective

- *So*, assuming $t \sim \mathcal{N}$, ML estimation leads to sum-of-squared errors minimisation.
- *Equivalently*, minimising sum-of-squared errors implies $t \sim \mathcal{N}$ (*i.e.*, noise was normally distributed).

Polynomial Curve Fitting
A Probabilistic Perspective

- $\mathbf{w}_{ML}$ and β_{ML} yields a probability distribution over the prediction t.

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}_{ML}, \beta_{ML}) = \prod_{n=1}^{N} \mathcal{N}(t_n | y(x_n, \mathbf{w}_{ML}), \beta_{ML}^{-1})$$

- The polynomial function $y(x, \mathbf{w}_{ML})$ alone only gives a point estimate of t.