CS-567 Machine Learning

Nazar Khan

PUCIT

Lecture 11
The Gaussian Distribution

The Gaussian Distribution

- The Gaussian distribution for a continuous, multivariate D-dimensional vector $\mathbf{x}$ is given by

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

  where the $D \times D$ matrix $\mathbf{\Sigma}$ is called the **covariance matrix** and $|\mathbf{\Sigma}|$ is its determinant.
- Gaussian distribution is intrinsically uni-modal. Its mode is the same as its mean $\boldsymbol{\mu}$.
- Cannot represent multi-modal data. For that a *mixture of Gaussians* can be used.

Mahalanobis Distance

- The term within the exponent is the so-called *squared-Mahalanobis distance*

$$d(\mathbf{x})^2 = (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

- All $\mathbf{x}$ satisfying $d(\mathbf{x}) = k$ constitute the k-th *iso-surface* of function $d(\cdot)$.
- Iso-surfaces of Mahalanobis distance are iso-surfaces of the Gaussian density also.

Σ – The Covariance Matrix

- Covariance matrix $\boldsymbol{\Sigma}$ is
  - Real-valued
  - Symmetric
  - Positive Definite (all eigenvalues are positive)
- Its eigen-decomposition can be written as

$$\boldsymbol{\Sigma} = \sum_{i=1}^{D} \lambda_i \mathbf{u}_i \mathbf{u}_i^T$$

- Using this eigen-decomposition, its inverse can be written as

$$\boldsymbol{\Sigma}^{-1} = \sum_{i=1}^{D} \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^T$$

Σ – The Covariance Matrix

- The eigen-decomposition of $\mathbf{\Sigma}^{-1}$ can be substituted in the squared-Mahalanobis distance

$$d(\mathbf{x})^2 = (\mathbf{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = (\mathbf{x}-\boldsymbol{\mu})^T\sum_{i=1}^{D}\frac{1}{\lambda_i}\mathbf{u}_i\mathbf{u}_i^T(\mathbf{x}-\boldsymbol{\mu})$$

$$= \sum_{i=1}^{D}\frac{1}{\lambda_i}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{u}_i\mathbf{u}_i^T(\mathbf{x}-\boldsymbol{\mu}) = \sum_{i=1}^{D}\frac{((\mathbf{x}-\boldsymbol{\mu})^T\mathbf{u}_i)^2}{\lambda_i}$$

- Projection of $\mathbf{x}-\boldsymbol{\mu}$ onto orthonormal basis $\mathbf{u}_1,\dots,\mathbf{u}_D$.
- Each projection onto $\mathbf{u}_i$ is divided by the variance λ_i along direction $\mathbf{u}_i$.
- Generalization of univariate Gaussian where exponent was $\frac{(x-\mu)^2}{\sigma^2}$. Now exponent is sum of $\frac{(\mathbf{x}^T\mathbf{u}_i-\boldsymbol{\mu}^T\mathbf{u}_i)^2}{\lambda_i}$.

Σ – The Covariance Matrix

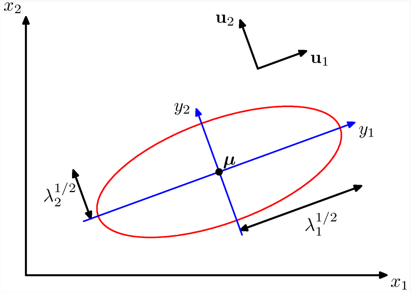
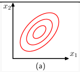
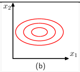
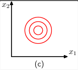


Figure: Elliptical iso-contour of a 2D Gaussian. Center of ellipse is determined by $\boldsymbol{\mu}$, axes are determined by the eigenvectors of $\boldsymbol{\Sigma}$ and axes lengths are determined via the eigenvalues of $\boldsymbol{\Sigma}$.

Σ – The Covariance Matrix

- Covariance matrix $\boldsymbol{\Sigma}$ can be categorised as

Category	$\boldsymbol{\Sigma}$ $(D=2)$	**DoF**	**Iso-contours** $(D=2)$
General	$\begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2 \\ \sigma_2\sigma_1 & \sigma_2^2 \end{pmatrix}$	$\frac{D(D+1)}{2}$	 (a)
Diagonal	$\begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$	D	 (b)
Isotropic	$\sigma^2\mathbf{I}$	1	 (c)

- Diagonal and isotropic cases are easy to work with but cannot represent data with interesting correlations.

Central Limit Theorem

- For random variables $\mathbf{x}_1, \ldots, \mathbf{x}_N$ that belong to any distribution (non-Gaussian), the sum $\mathbf{s} = \mathbf{x}_1 + \cdots + \mathbf{x}_N$ approaches a Gaussian random variable as N approaches ∞.
- This is known as the *Central Limit Theorem*.
- This is one reason for the popularity of the Gaussian distribution.
- Lots of natural phenomena correspond to sums or averages of many (non-Gaussian) random variables.
- For large enough N, these phenomena can be modelled by Gaussian distributions.

Fitting Gaussian density to data

- We have already covered how ML and MAP estimates for Gaussian density can be obtained.
- For computing log-likelihood of Gaussian, it is sufficient to pre-compute the following 2 statistics from the data:
  - the $D \times 1$ vector $\sum_{n=1}^{N} \mathbf{x}_n$
  - the $D \times D$ matrix $\sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^T$
- These statistics are called *sufficient statistics* for log-likelihood of Gaussian. The individual data items can be discarded once these are computed.