National Technical University of Ukraine
“Igor Sikorsky Kyiv Polytechnic Institute”
Institute of Physics and Technology

Lecture 6
Generative Models for Discrete Data. Gaussian Models

Dmytro Progonov,

PhD, Associate Professor,

Department Of Physics and Information Security Systems

Content

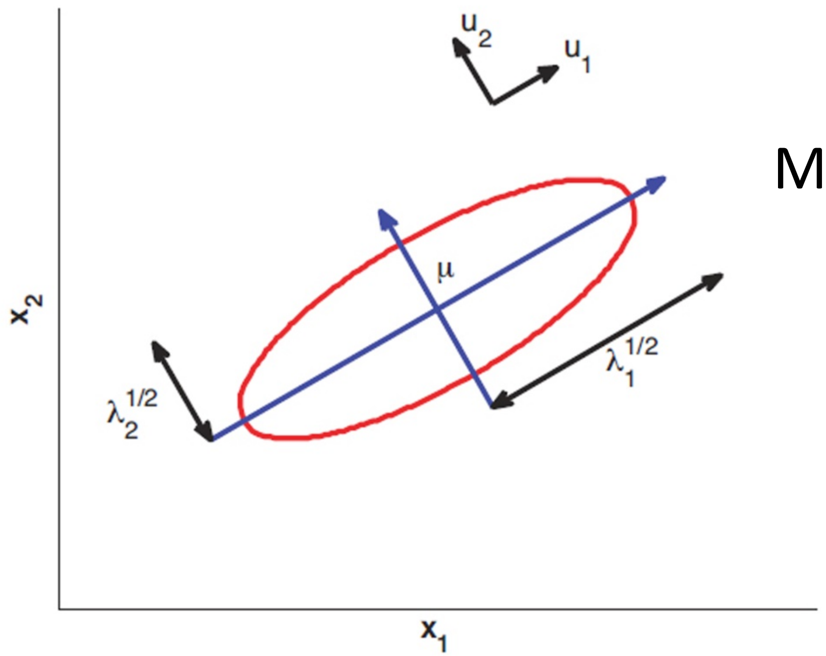
- Gaussian models definitions;
- Maximum Likelihood Estimation for MVN;
- Gaussian Discriminant Analysis;
- MLE for Discriminant Analysis;
- Strategies for preventing overfitting;
- Interference in jointly Gaussian distribution;
- Linear Gaussian systems.

Gaussian models definitions

Probability density function for an multivariate normal (MVN) distribution (MVN) in D dimensions is defined as:

$$\boxed{\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) \triangleq \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{D/2}}\exp\underbrace{\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]}_{}}$$

Mahalanobis distance between a data vector $\mathbf{x}$ and the mean vector $\boldsymbol{\mu}$



Visualization of a 2 dimensional Gaussian density. The major and minor axes of the ellipse are defined by the first two eigenvectors of the covariance matrix, namely $\mathbf{u}_1$ and $\mathbf{u}_2$. Based on Figure 2.7 of (Bishop 2006a).

Maximum Likelihood Estimation for MVN

<u>*Theorem (MLE for Gaussian)*</u>: if we have N iid samples $x_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the MLE for parameters is given by:

$$\boxed{\hat{\boldsymbol{\mu}}_{mle} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i \triangleq \bar{\mathbf{x}},}$$

$$\boxed{\hat{\boldsymbol{\Sigma}}_{mle} = \frac{1}{N}\sum_{i=1}^{N} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T = \frac{1}{N}\sum_{i=1}^{N} (\mathbf{x}_i {\mathbf{x}_i}^T) - \bar{\mathbf{x}}\bar{\mathbf{x}}^T.}$$

That is, the MLE is just the empirical mean and empirical covariance.

For univariate case, we get the following familiar results:

$$\boxed{\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i \triangleq \bar{x},}$$

$$\boxed{\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \bar{x})^2 = \frac{1}{N}\sum_{i=1}^{N} ({x_i}^2) - \bar{x}^2.}$$

Gaussian Discriminant Analysis (1/3)

One important application of MVN is to define the class conditional density in a generative classifier:

$$\boxed{p(\mathbf{x}|y=c,\boldsymbol{\theta}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_c,\boldsymbol{\Sigma}_c)}$$

We can classify the feature vector using the following decision rule (**nearest centroid classifier**):

$$\boxed{\hat{y}(\mathbf{x}) = \operatorname*{argmax}_c \left[\log\left(p(y=c|\boldsymbol{\pi})\right) + \log\left(p(\mathbf{x}|\boldsymbol{\theta}_c)\right)\right]}$$

Gaussian Discriminant Analysis (2/3)

By plugging in the definition of Gaussian density the posteriors over the class labels, we obtain **quadratic discriminant analysis**:

$$\boxed{p(\mathbf{x}|y=c,\boldsymbol{\theta}) = \frac{\pi_c|2\pi\boldsymbol{\Sigma}_c|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_c)^T{\boldsymbol{\Sigma}_c}^{-1}(\mathbf{x}-\boldsymbol{\mu}_c)\right]}{\sum_{\acute{c}}\pi_{\acute{c}}|2\pi\boldsymbol{\Sigma}_{\acute{c}}|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_{\acute{c}})^T{\boldsymbol{\Sigma}_{\acute{c}}}^{-1}(\mathbf{x}-\boldsymbol{\mu}_{\acute{c}})\right]}}$$

Consider a special case in which the covariance matrices are tied or shared across classes ($\boldsymbol{\Sigma}_c = \boldsymbol{\Sigma}$):

$$\boxed{p(\mathbf{x}|y=c,\boldsymbol{\theta}) \propto \exp\left[{\boldsymbol{\mu}_c}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}{\boldsymbol{\mu}_c}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_c + \log[\pi_c]\right]\exp\left[-\frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}\right]}$$

Let us define:

$$\boxed{\gamma_c = -\frac{1}{2}{\boldsymbol{\mu}_c}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_c + \log[\pi_c]} \qquad \boxed{\boldsymbol{\beta}_c = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_c}$$

Gaussian Discriminant Analysis (3/3)

Then we can write:

$$\boxed{p(\mathbf{x}|y=c,\boldsymbol{\theta}) = \frac{\exp\left[{\boldsymbol{\beta}_c}^T\mathbf{x}+\gamma_c\right]}{\sum_{\acute{c}}\exp\left[{\boldsymbol{\beta}_{\acute{c}}}^T\mathbf{x}+\gamma_{\acute{c}}\right]} = \mathcal{S}(\boldsymbol{\eta})_c}$$

where $\boldsymbol{\eta} = \left[{\boldsymbol{\beta}_1}^T\mathbf{x}+\gamma_1,\cdots,{\boldsymbol{\beta}_C}^T\mathbf{x}+\gamma_C\right]$, and $\mathcal{S}$ is softmax function defined as:

$$\boxed{\mathcal{S}(\boldsymbol{\eta})_c = \frac{e^{\eta_c}}{\sum_{\acute{c}=1}^{C} e^{\eta_{\acute{c}}}}}$$

If we take logs, we end up with linear function of $\mathbf{x}$ (because $\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}$) cancels from numerator/denominator). Thus the decision boundary between any two classes well be a straight line. Hence this technique is called **linear discriminant analysis (LDA)**.

MLE for Discriminant Analysis

The simplest way to fit a discriminant analysis model I to use maximum likelihood:

$$\boxed{\log[p(\mathcal{D}|\boldsymbol{\theta})] = \left[\sum_{i=1}^{N}\sum_{c-1}^{C}\mathbb{I}(y_i = c)\log[\pi_c]\right] + \sum_{c=1}^{C}\left[\sum_{i:y_i=c}\log[\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)]\right]}$$

We see that this factorizes into a term for $\boldsymbol{\pi}$ and C for each $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$. For the class conditional densities, we just partition the data based on its class label, and compute MLE for each Gaussian:

$$\boxed{\hat{\boldsymbol{\mu}}_c = \frac{1}{N_c}\sum_{i:y_i=c}\mathbf{x}_i\,,}$$

$$\boxed{\hat{\boldsymbol{\Sigma}}_c = \frac{1}{N_c}\sum_{i:y_i=c}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_c)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_c)^T\,.}$$

Strategies for preventing overfitting

- Use a diagonal covariance matrix for each class, which assumes the feature are conditionally independent; this is equivalent to using a naïve Bayes classifier;
- Use a full covariance matrix, but force it to be the same for all classes ($\boldsymbol{\Sigma}_c = \boldsymbol{\Sigma}$). This is an example of **parameter sharing**;
- Use a diagonal covariance matrix *and* force it to be shared. This is called **diagonal covariance LDA**;
- Use a full covariance matrix, but impose a prior and then integrate it out;
- Fit a full or diagonal covariance matrix by MAP estimation;
- Project the data into a low-dimensional subspace and fit the Gaussian here.

Interference in jointly Gaussian distribution

Given a join distribution $p(\mathbf{x}_1, \mathbf{x}_2)$ it is useful to be able to compute marginals $p(\mathbf{x}_1)$ and conditionals $p(\mathbf{x}_1|\mathbf{x}_2)$.

Theorem (marginals and conditionals for MVN): Suppose $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$ is jointly Gaussian with parameters

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}, \boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-\mathbf{1}} = \begin{pmatrix} \boldsymbol{\Lambda}_{11} & \boldsymbol{\Lambda}_{12} \\ \boldsymbol{\Lambda}_{21} & \boldsymbol{\Lambda}_{22} \end{pmatrix}$$

Then the marginals are given by:

$$\boxed{p(\mathbf{x}_1) = \mathcal{N}(\mathbf{x}_1|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})} \qquad \boxed{p(\mathbf{x}_2) = \mathcal{N}(\mathbf{x}_2|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})}$$

and the posterior conditional is given by:

$$\boxed{p(\mathbf{x}_1|\mathbf{x}_2) = \mathcal{N}\left(\mathbf{x}_1|\boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{1|2}\right)}$$

$$\boxed{\boldsymbol{\mu}_{1|2} = \boldsymbol{\Sigma}_{1|2}(\boldsymbol{\Lambda}_{11}\boldsymbol{\mu}_1 - \boldsymbol{\Lambda}_{12}(\mathbf{x}_2 - \boldsymbol{\mu}_2))} \qquad \boxed{\boldsymbol{\Sigma}_{1|2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}{}^{-1}\boldsymbol{\Sigma}_{21} = \boldsymbol{\Lambda}_{11}{}^{-\mathbf{1}}}$$

Linear Gaussian systems

Let $\mathbf{x} \in \mathbb{R}^{D_x}$ be a hidden variable and $\mathbf{y} \in \mathbb{R}^{D_y}$ be a noisy observation of $\mathbf{x}$. Let us assume we have the following prior and likelihood

$$\boxed{p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)} \qquad \boxed{p(\mathbf{y}|\mathbf{x}) = \mathcal{N}\left(\mathbf{y}|\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}, \boldsymbol{\Sigma}_y\right)}$$

where $\boldsymbol{A}$ is a matrix of size $D_y \times D_x$.

<u>Theorem (Bayes rule for linear Gaussian systems)</u>: Given a linear Gaussian system, the posterior $p(\mathbf{x}|\mathbf{y})$ is given by the following:

$$\boxed{p(\mathbf{x}|\mathbf{y}) = \mathcal{N}\left(\mathbf{x}|\boldsymbol{\mu}_{x|y}, \boldsymbol{\Sigma}_{x|y}\right)} \qquad \boxed{\boldsymbol{\mu}_{x|y} = \boldsymbol{\Sigma}_{x|y}\left(\boldsymbol{A}^T\boldsymbol{\Sigma}_y{}^{-1}(\boldsymbol{y} - \boldsymbol{b}) + \boldsymbol{\Sigma}_x{}^{-1}\boldsymbol{\mu}_x\right)}$$

$$\boxed{\boldsymbol{\Sigma}_{x|y}{}^{-1} = \boldsymbol{\Sigma}_x{}^{-1} + \boldsymbol{A}^T\boldsymbol{\Sigma}_y{}^{-1}\boldsymbol{A}}$$

The normalization constant $p(\mathbf{y})$ is given by:

$$\boxed{p(\mathbf{y}) = \mathcal{N}\left(\mathbf{y}|\boldsymbol{A}\boldsymbol{\mu}_x + \boldsymbol{b}, \boldsymbol{\Sigma}_y + \boldsymbol{A}^T\boldsymbol{\Sigma}_x\boldsymbol{A}\right)}$$

Conclusion

- Definitions of Gaussian models and Linear Gaussian systems were considered;
- Maximum Likelihood Estimation procedure for Gaussian Discriminant Analysis was introduced;
- Strategies for preventing overfitting were presented;
- Methods for interference in jointly Gaussian distribution was shown.