COURSE 02457

Signal Processing in Non-linear Systems:

Lecture 3

- Learning problem
- The likelihood function and Bayesian learning
- Maximum likelihood estimation of parameters in the normal distribution
- Learning by least squares
- Linear models
- Linear discriminants
- Logistic regression
- Fishers linear discriminant

The learning problem

- Supervised learning: Learning relations between sets of variables e.g. between input and output variables, conditional distributions $p(\text{output}|\text{input})$.
- Unsupervised learning: Learning the distribution of a set of variables $p(\text{input})$.

The Bayesian paradigm

- The density of the measured signals is modelled by a parameterized density: $p(x) \sim p(x|\theta)$.
- Let $\chi = \{x_1, x_2, x_3, ..., x_N\}$ be a $training\ set$
- Objective: Find the distribution of the parameter vector, $p(\theta|\chi)$, hence the parameters are considered stochastic.

The likelihood function

- Let $\chi = \{x_1, x_2, x_3, ..., x_N\}$ be a $training\ set$
- We use Bayes theorem

$$p(\theta|\chi) = \frac{p(\chi|\theta)p(\theta)}{p(\chi)}$$

- The function $p(\chi|\theta)$ is called the likelihood function (more correct the likelihood of the parameter vector θ). The density $p(\theta)$ is called the $a\ priori$ or $prior$ parameter distribution.
- If the prior is "flat" in the neighborhood of the peak of $p(\chi|\theta)$, we have

$$p(\theta|\chi) \propto p(\chi|\theta)$$

- ...and finding the most probable parameters is equivalent to finding the maximum likelihood parameters.

Maximum likelihood & optimization

- For independent examples, $\chi = \{x_1, x_2, x_3, ..., x_N\}$, the likelihood function factorize

$$p(\chi|\theta) = \prod_{n=1}^{N} p(x_n|\theta)$$

- Many algorithms are based on minimizing an index or costfunction

$$E(\theta) = -\log p(\chi|\theta) = \sum_{n=1}^{N} -\log p(x_n|\theta)$$

1D normal distribution

- Let the parameterized density be a 1D normal distribution

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

- For independent examples, $\chi = \{x_1, x_2, x_3, ..., x_N\}$, the likelihood function becomes

$$\begin{aligned} p(\chi|\mu,\sigma^2) &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left(-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\right) \\ E(\mu,\sigma^2) &= \frac{N}{2}\log 2\pi\sigma^2 + \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 \end{aligned}$$

1D normal distribution

- Costfunction for maximum likelihood estimation of mean and variance

$$E(\mu, \sigma^2) = \frac{N}{2} \log 2\pi\sigma^2 + \frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2$$

- Derivatives are zero as minimum:

$$\begin{aligned}
\frac{\partial E(\mu, \sigma^2)}{\partial \mu} &= \frac{1}{\sigma^2} \sum_{n=1}^{N} -(x_n - \mu) \\
\frac{\partial E(\mu, \sigma^2)}{\partial \sigma^2} &= \frac{N}{2} \frac{1}{\sigma^2} - \frac{1}{2(\sigma^2)^2} \sum_{n=1}^{N} (x_n - \mu)^2 \\
0 &= \frac{1}{\widehat{\sigma^2}} \sum_{n=1}^{N} -(x_n - \widehat{\mu}) \\
0 &= \frac{N}{2} \frac{1}{\widehat{\sigma^2}} - \frac{1}{2(\widehat{\sigma^2})^2} \sum_{n=1}^{N} (x_n - \widehat{\mu})^2 \\
\widehat{\mu} &= \frac{1}{N} \sum_{n=1}^{N} x_n \\
\widehat{\sigma^2} &= \frac{1}{N} \sum_{n=1}^{N} (x_n - \widehat{\mu})^2
\end{aligned}$$

Multivariate normal distribution

- For independent examples, $\chi = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, ..., \mathbf{x}_N\}$, the likelihood function becomes

$$p(\chi|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \left(\frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}}\right)^N \exp\left(-\frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu})\right)$$

$$E(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{N}{2}\log|2\pi\boldsymbol{\Sigma}| + \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu})$$

- We need two rules:

$$\frac{\partial \log|\mathbf{A}|}{\partial u} = \mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial u}$$

$$\frac{\partial \mathbf{A}^{-1}}{\partial u} = -\mathbf{A}^{-1}\frac{\partial \mathbf{A}}{\partial u}\mathbf{A}^{-1}$$

Multivariate normal distribution

$$
\begin{aligned}
\frac{\partial E(\boldsymbol{\mu}, \sigma^2)}{\partial \boldsymbol{\mu}} &= \sum_{n=1}^{N} -(\mathbf{x}_n - \boldsymbol{\mu})\boldsymbol{\Sigma}^{-1} \\
\frac{\partial E(\boldsymbol{\mu}, \sigma^2)}{\partial \boldsymbol{\Sigma}} &= \frac{N}{2}\boldsymbol{\Sigma}^{-1} - \frac{1}{2}\boldsymbol{\Sigma}^{-1}\left(\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})(\mathbf{x}_n - \boldsymbol{\mu})'\right)\boldsymbol{\Sigma}^{-1} \\
0 &= \sum_{n=1}^{N} -(\mathbf{x}_n - \widehat{\boldsymbol{\mu}})\widehat{\boldsymbol{\Sigma}}^{-1} \\
0 &= \frac{N}{2}\widehat{\boldsymbol{\Sigma}}^{-1} - \frac{1}{2}\widehat{\boldsymbol{\Sigma}}^{-1}\sum_{n=1}^{N}\left((\mathbf{x}_n - \widehat{\boldsymbol{\mu}})(\mathbf{x}_n - \widehat{\boldsymbol{\mu}})'\right)\widehat{\boldsymbol{\Sigma}}^{-1} \\
\widehat{\boldsymbol{\mu}} &= \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n \\
\widehat{\boldsymbol{\Sigma}} &= \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \widehat{\boldsymbol{\mu}})(\mathbf{x}_n - \widehat{\boldsymbol{\mu}})'
\end{aligned}
$$

Least squares as maximum likelihood

- Let $(\chi_{\mathbf{x}}, \chi_y) = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), (\mathbf{x}_3, y_3), ..., (\mathbf{x}_N, y_N)\}$,
- We seek a conditional density model of the form

$$
\begin{aligned}
y &= f_\theta(\mathbf{x}) + \nu \\
p(y|\mathbf{x}, \sigma^2, \theta) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y - f_\theta(\mathbf{x}))^2\right) \\
p(\chi_y|\chi_{\mathbf{x}}, \sigma^2, \theta) &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left(-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - f_\theta(\mathbf{x}_n)^2\right) \\
E(\theta, \sigma^2) &= \frac{N}{2}\log 2\pi\sigma^2 + \frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - f_\theta(\mathbf{x}_n))^2
\end{aligned}
$$

- Hence, maximizing the likelihood for Gaussian noise leads to a least squares problem (for θ).

Discriminant functions

- A signal detection system divides signal/measurement space in regions $\mathcal{R}$. A set of $discriminant\ functions$ $y_j(\mathbf{x})$ are defined so that

$$y_j(\mathbf{x}) > y_k(\mathbf{x}) \quad j \neq k, \mathbf{x} \in \mathcal{R}_j$$

- Bayes decision theory:

$$y_k(\mathbf{x}) = P(C_k|\mathbf{x})$$

- Special case for binary decisions: A single function defines the decision boundary:

$$y_(\mathbf{x}) = y_1(\mathbf{x}) - y_2(\mathbf{x}) = 0$$

The linear model

- Linear discriminant function for two classes

$$y(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0$$

- Terminology: $\mathbf{w}$ are called the weights, and w_{k0} is called the threshold.
- Simplify by dummy input

$$y(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0 = \tilde{\mathbf{w}}^T\tilde{\mathbf{x}}$$

- $\tilde{\mathbf{w}}^T = (w_0, \mathbf{w})$ and $\tilde{\mathbf{x}} = (1, \mathbf{x})$

The linear discriminant

- Linear discriminant functions for multiple classes

$$y_k(\mathbf{x}) = \mathbf{w}_k^T \mathbf{x} + w_{k0}$$

- Deciding between two classes j, k

$$y_k(\mathbf{x}) - y_j(\mathbf{x}) = (\mathbf{w}_k - \mathbf{w}_j)^T \mathbf{x} + (w_{k0} - w_{j0})$$

- Decision boundary between two classes j, k

$$(\mathbf{w}_k - \mathbf{w}_j)^T \mathbf{x} + (w_{k0} - w_{j0}) = 0$$

Decision regions

- Decision regions of the multiclass linear discriminant are convex (and simply connected)

$$\widehat{\mathbf{x}} = \alpha \mathbf{x}^A + (1-\alpha)\mathbf{x}^B$$

- Let $\mathbf{x}^A, \mathbf{x}^B \in \mathcal{R}_k$, hence $y_k(\mathbf{x}^A) > y_j(\mathbf{x}^A)$ and $y_k(\mathbf{x}^B) > y_j(\mathbf{x}^B)$.

$$\begin{aligned} y_k(\widehat{\mathbf{x}}) &= \mathbf{w}_k^T(\alpha \mathbf{x}^A + (1-\alpha)\mathbf{x}^B) \\ &= \alpha y_k(\mathbf{x}^A) + (1-\alpha) y_k(\mathbf{x}^B) \\ &> \alpha y_j(\mathbf{x}^A) + (1-\alpha) y_j(\mathbf{x}^B) \\ &= \alpha \mathbf{w}_j^T \mathbf{x}^A + (1-\alpha)\mathbf{w}_j^T \mathbf{x}^B \\ &= \mathbf{w}_j^T(\alpha \mathbf{x}^A + (1-\alpha)\mathbf{x}^B) \\ &= y_j(\widehat{\mathbf{x}}) \end{aligned}$$

- Thus all points along the line between $\mathbf{x}^A$ and $\mathbf{x}^B$ are contained in the decision region $\mathcal{R}_k$ (convex and simply connected).

Logistic regression

- Let the class-conditional probability densities for a two-class problem be given by

$$p(\mathbf{x}|C_k) = \frac{1}{\sqrt{(2\pi)^d|\mathbf{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right)$$

- where the classes have identical covariance matrices $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{\Sigma}$
- In this case the posterior probabilities are

$$\begin{aligned} p(C_1|\mathbf{x}) &= \frac{p(\mathbf{x}|C_1)P(C_1)}{p(\mathbf{x}|C_1)P(C_1) + p(\mathbf{x}|C_2)P(C_2)} \\ p(C_2|\mathbf{x}) &= \frac{p(\mathbf{x}|C_2)P(C_2)}{p(\mathbf{x}|C_1)P(C_1) + p(\mathbf{x}|C_2)P(C_2)} \end{aligned}$$

$$\begin{aligned} p(C_1|\mathbf{x}) &= \frac{1}{1 + p(\mathbf{x}|C_2)P(C_2)/p(\mathbf{x}|C_1)P(C_1)} \\ &= \frac{1}{1 + \exp\left(-a(\mathbf{x})\right)} \end{aligned}$$

Logistic regression cont'd

- The logistic regression Bayes decisions are based on

$$p(C_1|\mathbf{x}) \;=\; \frac{1}{1+\exp\left(-a(\mathbf{x})\right)}$$

- $p(C_1|\mathbf{x}) > 0.5$ when the linear discriminant function given by

$$\begin{aligned} a(\mathbf{x}) \;&=\; \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2) \\ &-\; \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) + \log\frac{P(C_2)}{P(C_1)} \\ &=\; (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 \\ &\quad + \log\frac{P(C_2)}{P(C_1)} \end{aligned}$$

  ... is positive

Logistic regression cont'd

- Hence, we have a recipy for designing a two class detector:
  Estimate the two class mean vectors and the common covariance matrix

$$
\begin{aligned}
\boldsymbol{\mu}_1 &= \frac{1}{N_1}\sum_{n=1}^{N_1}\mathbf{x}^n \\
\boldsymbol{\mu}_2 &= \frac{1}{N_2}\sum_{n=1}^{N_2}\mathbf{x}^n \\
\boldsymbol{\Sigma} &= \frac{1}{N_1+N_2}\left(\sum_{n=1}^{N_1}(\mathbf{x}^n-\boldsymbol{\mu}_1)(\mathbf{x}^n-\boldsymbol{\mu}_1)^T\right. \\
&\quad \left.+\sum_{n=1}^{N_2}(\mathbf{x}^n-\boldsymbol{\mu}_2)(\mathbf{x}^n-\boldsymbol{\mu}_2)^T\right) \\
P(C_1) &= \frac{N_1}{N_1+N_2} \\
P(C_2) &= \frac{N_2}{N_1+N_2}
\end{aligned}
$$

Least squares techniques

- Let a training set be given by $\mathcal{D} = \{(t^1, \mathbf{x}^1), ..., (t^N, \mathbf{x}^N)\}$, the sum-of-squares approximation error is given by

$$E = \frac{1}{2} \sum_{n=1}^{N} (\mathbf{w}^T \mathbf{x}^n + w_0 - t^n)^2 \qquad (1)$$

- The optimal parameters are found by gradient based minimization,

$$\frac{\partial E}{\partial \mathbf{w}} = \sum_{n=1}^{N} (\mathbf{w}^T \mathbf{x}^n + w_0 - t^n) \mathbf{x}^n$$

$$\frac{\partial E}{\partial w_0} = \sum_{n=1}^{N} (\mathbf{w}^T \mathbf{x}^n + w_0 - t^n)$$

Least squares techniques cont'd

- equations to solve

$$\sum_{n=1}^{N}(\mathbf{w}^T\mathbf{x}^n + w_0 - t^n)\mathbf{x}^n \;=\; 0$$

$$\sum_{n=1}^{N}(\mathbf{w}^T\mathbf{x}^n + w_0 - t^n) \;=\; 0$$

- the solution is given by in terms of $\boldsymbol{\mu} = (1/N)\sum \mathbf{x}^n$, and $\tau = (1/N)\sum t^n$

$$\mathbf{w} \;=\; -\left(\frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}^n - \boldsymbol{\mu})(\mathbf{x}^n - \boldsymbol{\mu})^T\right)^{-1}\left(\frac{1}{N}\sum_{n=1}^{N}(t^n - \tau)\mathbf{x}^n\right)$$

$$w_0 \;=\; -\mathbf{w}^T\boldsymbol{\mu} + \tau$$

- This can be used to model any linear input-output relation

Fishers linear discriminant

- Specific encoding $t_+^n = N/N_1$, $t_-^n = -N/N_2$

$$\mathbf{w} \propto \boldsymbol{S}_W^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$$

- where $\boldsymbol{S}_W \equiv \boldsymbol{\Sigma}_{\text{logistic}}$

$$\begin{aligned}\boldsymbol{S}_W &= \frac{1}{N}\sum_{n=1}^{N_1}(\mathbf{x}^n - \boldsymbol{\mu}_1)(\mathbf{x}^n - \boldsymbol{\mu}_1)^T \\ &+ \frac{1}{N}\sum_{n=1}^{N_2}(\mathbf{x}^n - \boldsymbol{\mu}_2)(\mathbf{x}^n - \boldsymbol{\mu}_2)^T\end{aligned}$$

- Hence, same solution as for the logistic regression system aka "Fishers linear discriminant"