COURSE 02457

Signal Processing in Non-linear Systems:

Lecture 2

- Probabilities and densities
- Conditional probabilities and densities
- Bayes' theorem
- 1D Normal distribution
- Multivariate normal distribution
- Correlations
- Signal detection with normal distributions
- Features & the curse of dimensionality

Conditional probabilities, Bayes' theorem

Figure 1: The measured signals take (discrete) values X and each signal is assigned to one of the classes $\mathcal{C}_1, \mathcal{C}_2$. The number of dots in each celle corresponds to the number of signals that fall in the given class and have taken on the value X

$$
\begin{aligned}
P(\mathcal{C}_k, X^l) &= P(X^l|\mathcal{C}_k)P(\mathcal{C}_k) \\
P(\mathcal{C}_k, X^l) &= P(\mathcal{C}_k|X^l)P(X^l)
\end{aligned}
$$

$$
\begin{aligned}
P(\mathcal{C}_k|X^l) &= \frac{P(X^l|\mathcal{C}_k)P(\mathcal{C}_k)}{P(X^l)} \\
P(X^l|\mathcal{C}_k) &= \frac{P(\mathcal{C}_k|X^l)P(X^l)}{P(\mathcal{C}_k)}
\end{aligned}
$$

Conditional probabilities cont'd

$$
\begin{aligned}
P(\mathcal{C}_1, |X^l) + P(\mathcal{C}_2|X^l) &= 1 \\
\frac{P(X^l|\mathcal{C}_1)P(\mathcal{C}_1)}{P(X^l)} + \frac{P(X^l|\mathcal{C}_2)P(\mathcal{C}_2)}{P(X^l)} &= 1 \\
P(X^l|\mathcal{C}_1)P(\mathcal{C}_1) + P(X^l|\mathcal{C}_2)P(\mathcal{C}_2) &= P(X_l)
\end{aligned}
$$

Conditional probabilities cont'd

Figure 2: Schematic plot of the histograms for a measured signal drawn from either of two populations $\mathcal{C}_1, \mathcal{C}_2$

Figure 3: Corresponding $P(X)$

Figure 4: Corresponding $P(\mathcal{C}|X)$'s

The probability density function $p(x)$

- In one dimension, the probability density function $p(x)$ is characterized by

$$P(x \in [a,b]) = \int_a^b p(x)dx$$

and expectations are computed by

$$\mathcal{E}(f(x)) = \int_{\text{Domain of x}} f(x)p(x)dx$$

the density function is normalized

$$P(x \in \text{Domainofx}) = \int_{\text{Domain of x}} p(x)dx = 1$$

The 'average value of x' (the mean of x)

$$\mathcal{E}(x) \equiv \mu = \int_{\text{Domain of x}} xp(x)dx$$

The spread of x around it's mean (the standard deviation)

$$\sigma = \sqrt{\int_{\text{Domain of x}} (x-\mu)^2 p(x)dx}$$

The probability density function $p(x)$

- In the multivariate case the probability density function $p(\mathbf{x})$ is characterized by

$$P(x_j \in [a_j, b_j] | j = 1, .., d) = \int_{a_1}^{b_1} ... \int_{a_d}^{b_d} p(\mathbf{x}) d\mathbf{x}$$

and expectations are computed by

$$\mathcal{E}(f(\mathbf{x})) = \int_{\text{Domain of } \mathbf{x}} f(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$

the density function is normalized $\mathcal{E}(1) = 1$.

The 'average value of $\mathbf{x}$' (the mean of $\mathbf{x}$)

$$\mathcal{E}(\mathbf{x}) \equiv \boldsymbol{\mu} = \int_{\text{Domain of } \mathbf{x}} \mathbf{x} p(\mathbf{x}) d\mathbf{x}$$

The spread of $\mathbf{x}$ around it's mean (the standard deviation) needs to be characterized by a matrix!

$$\boldsymbol{\Sigma} = \int_{\text{Domain of } \mathbf{x}} (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top p(\mathbf{x}) d\mathbf{x}$$

Bayes' theorem – multivariate version

$$
\begin{aligned}
P(\mathcal{C}_k, \mathbf{x}) &= p(\mathbf{x}|\mathcal{C}_k)P(\mathcal{C}_k) \\
P(\mathcal{C}_k, \mathbf{x}) &= P(\mathcal{C}_k|\mathbf{x})p(\mathbf{x})
\end{aligned}
$$

$$
\begin{aligned}
P(\mathcal{C}_k|\mathbf{x}) &= \frac{p(\mathbf{x}|\mathcal{C}_k)P(\mathcal{C}_k)}{p(\mathbf{x})} \\
p(\mathbf{x}|\mathcal{C}_k) &= \frac{P(\mathcal{C}_k|X^l)p(\mathbf{x})}{P(\mathcal{C}_k)}
\end{aligned}
$$

$$
\sum_{k=1}^{c} P(\mathcal{C}_k|\mathbf{x}) = 1
$$

$$
\sum_{k=1}^{c} p(\mathbf{x}|\mathcal{C}_k)P(\mathcal{C}_k) = p(\mathbf{x})
$$

Signal detection: Bayes decision theory

- A signal detection system (or pattern classifier) provides a rule for assigning a measurement to a given signal category (class)
- Hence, a classifier divides measument space (feature space) into disjoint regions $\mathcal{R}_1, \mathcal{R}_2, ..., \mathcal{R}_c$, such that measurements that fall into region $\mathcal{R}_k$ are assigned with class $\mathcal{C}_k$.
- Boundaries between regions are denoted decision surfaces or decision boundaries

Signal Detection: Bayes decision theory

Figure 5: Schematic plot of the densities for a measured signal drawn from either of two populations $\mathcal{C}_1, \mathcal{C}_2$

$$
\begin{aligned}
P(\text{error}) &= P(x \in \mathcal{R}_2, \mathcal{C}_1) + P(x \in \mathcal{R}_1, \mathcal{C}_2) \\
&= P(x \in \mathcal{R}_2 | \mathcal{C}_1) P(\mathcal{C}_1) + P(x \in \mathcal{R}_1 | \mathcal{C}_2) P(\mathcal{C}_2) \\
&= \left(\int_{\mathcal{R}_2} p(x|\mathcal{C}_1) dx \right) P(\mathcal{C}_1) + \left(\int_{\mathcal{R}_1} p(x|\mathcal{C}_2) dx \right) P(\mathcal{C}_2)
\end{aligned}
$$

- The probability of error is minimized if we assign points to $\mathcal{R}_1$, whenever $p(x|\mathcal{C}_1)P(\mathcal{C}_1) > p(x|\mathcal{C}_2)P(\mathcal{C}_2)$
- Using Bayes' theorem, this is equivalent to assign points to $\mathcal{R}_1$, whenever $p(\mathcal{C}_1|x) > p(\mathcal{C}_2|x)$, since we can divide by $p(\mathbf{x})$ on both sides of the inequality. Hence, the Bayes optimal signal detection system chooses the most probable class given the measurement.

The loss matrix

- The $loss$ matrix with elements $L_{k,j}$ specifies the penalty, or loss incured by assigning a signal to class $\mathcal{C}_j$ when it in fact belongs to class $\mathcal{C}_k$

- The expected loss for patterns in class k, and the total $risk$ are

$$\begin{aligned} R_k &= \sum_{j=1}^{c} L_{k,j} \int_{\mathcal{R}_j} p(\mathbf{x}|\mathcal{C}_k) d\mathbf{x} \\ R &= \sum_{k=1}^{c} R_k P(\mathcal{C}_k) \\ &= \sum_{j=1}^{c} \int_{\mathcal{R}_j} \sum_{k=1}^{c} L_{k,j} p(\mathbf{x}|\mathcal{C}_k) P(\mathcal{C}_k) d\mathbf{x} \end{aligned}$$

- The risk is minimized if the integrand is minimized in all points $\mathbf{x}$, hence, if the region $\mathcal{R}_j$ is chosen so that $\sum_{k=1}^{c} L_{k,j} p(\mathbf{x}|\mathcal{C}_k) P(\mathcal{C}_k) < \sum_{k=1}^{c} L_{k,i} p(\mathbf{x}|\mathcal{C}_k) P(\mathcal{C}_k)$

Uncertain decisions: Rejection mechanisms

- For signal detection systems for nosiy data (class overlap) we can reduce the error rate by rejecting uncertain decisions if the probability of correct classification is low

$$\max_k P(\mathcal{C}_k|\mathbf{x}) < \theta, \quad \theta \in [0, 1]$$

- The rejection threshold θ is a control parameter the $reject\ rate$ is given by:

$$\rho(\theta) = \int \Theta\left(\theta - \max_k P(\mathcal{C}_k|\mathbf{x})\right) p(\mathbf{x}) d\mathbf{x}$$

The uni-variate normal distribution

- In one dimension, the normal distribution's probability density function is given by

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

where the mean value parameter is

$$\mu = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx$$

and the variance,

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx$$

Multivariate normal distribution

- In d dimensions, the multivariate normal probability density function is given by

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

  where $\boldsymbol{\mu}$ is a d-dimensional vector, and $\boldsymbol{\Sigma}$ is a $d \times d$ covariance matrix.

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:

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

Detecting signals with normal distributions

- The k'th class has a prior probability $P(\mathcal{C}_k)$
  and is normally distributed:

$$p(\mathbf{x}|\mathcal{C}_k) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}_k|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)'\mathbf{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right)$$

- The Bayes rule is based on the posterior

$$P(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)P(\mathcal{C}_k)}{p(\mathbf{x})}$$

- or the *discriminant functions*

$$\begin{aligned} y_k(\mathbf{x}) &= \log p(\mathbf{x}|\mathcal{C}_k)P(\mathcal{C}_k) \\ &= -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)'\mathbf{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k) - \frac{1}{2}\log|\mathbf{\Sigma}_k| + \log P(\mathcal{C}_k) \end{aligned}$$

Detecting signals with normal distributions

- Special case: All signal components k have the same covariance matrix:

$$y_k(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_k) - \frac{1}{2}\log|\boldsymbol{\Sigma}| + \log P(\mathcal{C}_k)$$

$$\begin{aligned} y_k(\mathbf{x}) - y_j(\mathbf{x}) &= (\boldsymbol{\mu}_k - \boldsymbol{\mu}_j)'\Sigma^{-1}\mathbf{x} \\ &- \frac{1}{2}\boldsymbol{\mu}_k'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_j'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_j + \log\frac{P(\mathcal{C}_k)}{P(\mathcal{C}_j)} \end{aligned}$$

- NOTE: The decisions made in this case are based on linear decision surfaces

The curse of dimensionality

- To specify a map (e.g. a discriminant function) on a d-dimensional space by dividing the relevant parts of the this space into L cells pr. dimension requires L^d cells.

Features

- Often $correlations$ among variables means that there are only a few relevant dimensions. "Needle in a hay-stack" problems.
- Such relevant degrees of freedom are called $features$.
- Examples:
  - Image features for melanoma detection
  - Adaptive measures for digit recognition
  - Texture measures for glaucoma

Principal component analysis

- Objective to map $\mathbf{x} \rightarrow \mathbf{z}$ where the dimension of $\mathbf{z}$ is smaller than $\mathbf{x}$, without loosing signal variance.
- Let $\mathbf{u}_j$ be an orthonormal basis set in the space of $\mathbf{x}$

$$\begin{aligned}
\mathbf{x} &= \sum_{i=1}^{d} z_i \mathbf{u}_i \\
\delta_{i,j} &= \mathbf{u}_i^T \mathbf{u}_j \\
z_i &= \mathbf{u}_i^T \mathbf{x} \\
\tilde{\mathbf{x}} &= \sum_{i=1}^{M} z_i \mathbf{u}_i + \sum_{i=M+1}^{d} b_i \mathbf{u}_i
\end{aligned}$$

Principal component analysis cont'd

- The coefficients b_i are estimated using by the following argument for N example $\mathcal{D} = \{\mathbf{x}_\infty, \mathbf{x}_\in, ..., \mathbf{x}_\mathcal{N}\}$

$$
\begin{aligned}
\mathbf{x}^n - \tilde{\mathbf{x}^n} &= \sum_{i=M+1}^{d} (z_i^n - b_i)\mathbf{u}_i \\
E_M &= \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}^n - \tilde{\mathbf{x}}^n)^2 \\
&= \frac{1}{2}\sum_{n=1}^{N}\sum_{i=M+1}^{d} (z_i^n - b_i)^2 \\
b_i &= \frac{1}{N}\sum_{n=1}^{N} z_i^n \mathbf{u}_i^T \overline{\mathbf{x}} \\
\overline{\mathbf{x}} &= \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}^n
\end{aligned}
$$

Principal component analysis cont'd

- We can rewrite the sum-of-squares error,

$$
\begin{aligned}
E_M &= \frac{1}{2}\sum_{n=1}^{N}\sum_{i=M+1}^{d}\{\mathbf{u}_i^T(\mathbf{x}^n - \overline{\mathbf{x}})\}^2 \\
&= \frac{N}{2}\sum_{d=M+1}^{d}\mathbf{u}_i^T\mathbf{\Sigma}\mathbf{u}_i \\
\mathbf{\Sigma} &= \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}^n - \overline{\mathbf{x}})(\mathbf{x}^n - \overline{\mathbf{x}})^T
\end{aligned}
$$

- typo in equation (8.21) in Bishop, misses a factor $1/N$ in definition of $\mathbf{\Sigma}$

Principal component analysis cont'd

- Let $\mathbf{u}_i$ be the eigenvectors of $\mathbf{\Sigma}$, and λ_i the eigenvalues, then

$$E_M = \frac{N}{2} \sum_{d=M+1}^{d} \lambda_i$$

- Hence, we can minimize this error by choosing the eigenvalues to be the set of the $(d - (M + 1))$ smallest eigenvalues.
- The map we have sought is then from $\mathbf{x}$ to the vector $\mathbf{z} = (z_1, z_2, ..., z_M)$