Lecture 6: Bayes Classification

Richa Singh

Google classroom code: wgzuohn

Slides are prepared from several information sources including Duda, Hart, Stork

Recap: Bayes' Classification

- Posterior, likelihood, prior, evidence

$$\boxed{P(\omega_j|x) = \frac{p(x|\omega_j)P(\omega_j)}{p(x)},}$$

  – Evidence: In case of two categories

$$\boxed{p(x) = \sum_{j=1}^{2} p(x|\omega_j)P(\omega_j)}$$

$$\boxed{posterior = \frac{likelihood \times prior}{evidence}}$$

Bayesian Decision Theory

- Generalization of the preceding ideas
  - Use of more than one feature
  - Use more than two states of nature
  - Allowing actions other than decide on the state of nature
    - Allowing actions other than classification primarily allows the possibility of rejection
    - Refusing to make a decision in close or bad cases!
  - Introduce a loss function which is more general than the probability of error
    - The loss function states how costly each action taken is

Bayesian Decision Theory – Continuous Features…

- Let $\{\omega 1, \omega 2, \ldots, \omega c\}$ be the set of c states of nature (or "categories")
- Let $\{\alpha 1, \alpha 2, \ldots, \alpha a\}$ be the set of possible actions
- Let $\lambda(\alpha i \mid \omega j)$ be the loss incurred for taking action αi when the true state of nature is ωj

Two-category Classification

- $\alpha 1$: deciding $\omega 1$
- $\alpha 2$: deciding $\omega 2$
- $\lambda ij = \lambda(\alpha i \mid \omega j)$
- Loss incurred for deciding αi when the true state of nature is ωj

Two-category Classification

- $\alpha 1$: deciding $\omega 1$
- $\alpha 2$: deciding $\omega 2$
- $\lambda ij = \lambda(\alpha i \mid \omega j)$
- Loss incurred for deciding αi when the true state of nature is ωj
- Conditional risk:

$$
\begin{aligned}
R(\alpha_1|\mathbf{x}) &= \lambda_{11}P(\omega_1|\mathbf{x}) + \lambda_{12}P(\omega_2|\mathbf{x}) \\
R(\alpha_2|\mathbf{x}) &= \lambda_{21}P(\omega_1|\mathbf{x}) + \lambda_{22}P(\omega_2|\mathbf{x}).
\end{aligned}
$$

Two-category Classification

- Our rule is the following:

  if $R(\alpha 1 \mid x) < R(\alpha 2 \mid x)$
- Action $\alpha 1$: "decide $\omega 1$" is taken
- This results in the equivalent rule :
- Decide $\omega 1$ if:

$$(\lambda_{21} - \lambda_{11})p(\mathbf{x}|\omega_1)P(\omega_1) > (\lambda_{12} - \lambda_{22})p(\mathbf{x}|\omega_2)P(\omega_2)$$

- and decide $\omega 2$ otherwise

Bayesian Decision Theory – Continuous Features…

- Overall risk

  R = Sum of all $\underbrace{R(\alpha i \mid x)}_{\textbf{Conditional risk}}$ for i = 1,…,a

- Minimizing R ⟺ Minimizing R(αi | x) for i = 1,…, a

- $$R(\alpha_i \mid x) = \sum_{j=1}^{j=c} \lambda(\alpha_i \mid \omega_j) P(\omega_j \mid x)$$ for i = 1,…,a

Bayesian Decision Theory – Continuous Features…

- Select the action αi for which $R(\alpha i \mid x)$ is minimum

- R is minimum and is called the Bayes risk = best performance that can be achieved!

Likelihood Ratio

$$\boxed{(\lambda_{21} - \lambda_{11})p(\mathbf{x}|\omega_1)P(\omega_1) > (\lambda_{12} - \lambda_{22})p(\mathbf{x}|\omega_2)P(\omega_2)}$$

- The preceding rule is equivalent to the following rule:
- If $$\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} > \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \frac{P(\omega_2)}{P(\omega_1)}$$
- Then take action α1 (decide ω1)
- Otherwise take action α2 (decide ω2)

Likelihood Ratio…

- Optimal decision property

  “If the likelihood ratio exceeds a threshold value independent of the input pattern x, we can take optimal actions”

Questions?

The Normal Density

- Univariate density
  - Continuous density
  - A lot of processes are asymptotically Gaussian
  - Handwritten characters, speech sounds are ideal or prototype corrupted by random process (central limit theorem)

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

Univariate density

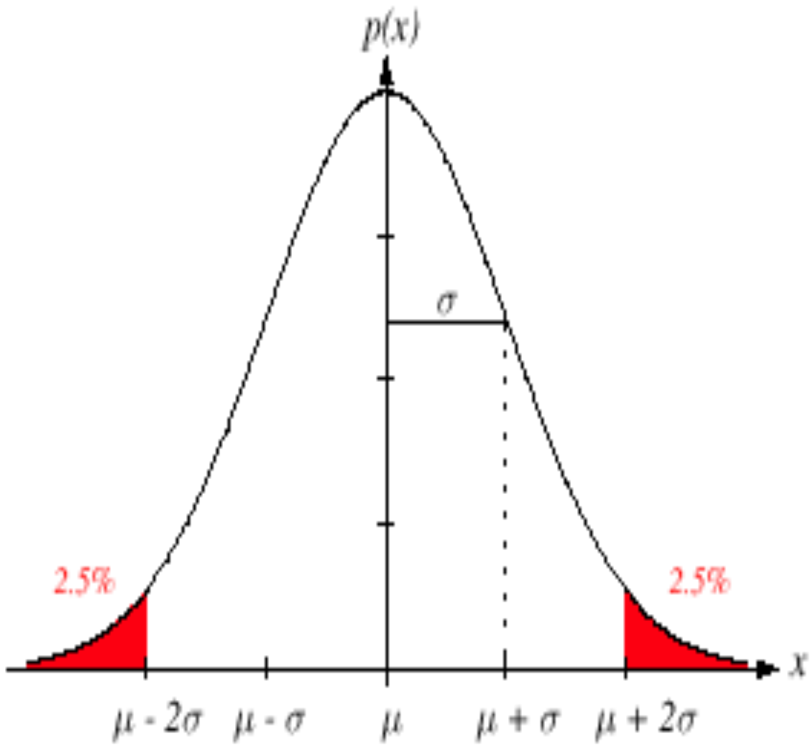

FIGURE 2.7. A univariate normal distribution has roughly 95% of its area in the range $|x - \mu| \leq 2\sigma$, as shown. The peak of the distribution has value $p(\mu) = 1/\sqrt{2\pi}\sigma$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

The Normal Density

- Multivariate density: Multivariate normal density in d dimensions is:

$$N(x;\mu,\sigma^2)=\frac{1}{(2\pi)^{d/2}\left|\Sigma\right|^{1/2}}\exp\left[-\frac{1}{2}(x-\mu)^t\Sigma^{-1}(x-\mu)\right]$$

$x = (x_1, x_2, \ldots, x_d)^t$

$\mu = (\mu_1, \mu_2, \ldots, \mu_d)^t$ mean vector

$\Sigma = d*d$ covariance matrix

$|\Sigma|$ and Σ^{-1} are determinant and inverse respectively

Multivariate density

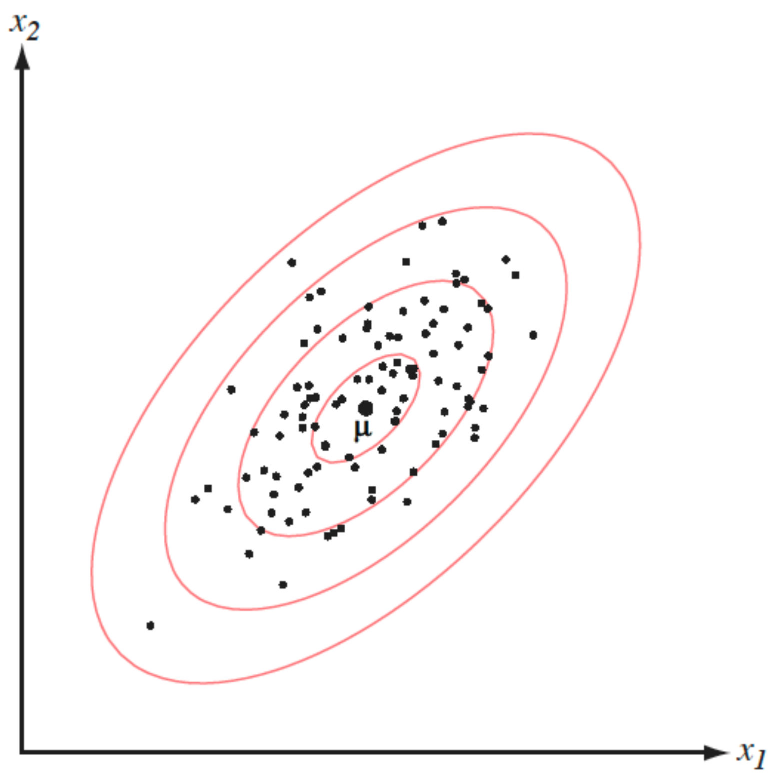

Figure 2.9: Samples drawn from a two-dimensional Gaussian lie in a cloud centered on the mean $\boldsymbol{\mu}$. The red ellipses show lines of equal probability density of the Gaussian.

Discriminant Functions for the Normal Density

$$P(\omega_j|x) = \frac{p(x|\omega_j)P(\omega_j)}{p(x)} \qquad P(error|x) = \min\,[P(\omega_1|x), P(\omega_2|x)]$$

- Minimum error-rate classification can be achieved by the discriminant function

- $g1(x) = \ln p(x \mid \omega 1) + \ln P(\omega 1) - \ln p(x)$
- $g2(x) = \ln p(x \mid \omega 2) + \ln P(\omega 2) - \ln p(x)$

Discriminant Functions for the Normal Density

$$P(\omega_j|x) = \frac{p(x|\omega_j)P(\omega_j)}{p(x)}$$

$$P(error|x) = \min\left[P(\omega_1|x), P(\omega_2|x)\right]$$

- Minimum error-rate classification can be achieved by the discriminant function

- $g1(x) = \ln p(x \mid \omega 1) + \ln P(\omega 1) - \ln p(x)$
- $g2(x) = \ln p(x \mid \omega 2) + \ln P(\omega 2) - \ln p(x)$

Discriminant Functions for the Normal Density

$$P(\omega_j|x) = \frac{p(x|\omega_j)P(\omega_j)}{p(x)},$$

- Minimum error-rate classification can be achieved by the discriminant function
- gi(x) = ln P(x | ωi) + ln P(ωi)

- Case of multivariate normal

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^t\Sigma_i^{-1}(x-\mu_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

Questions?