Bayesian Learning

Debrup Chakraborty

To be covered today

- Bayes decision theory
- Multivariate Normal Distribution
- Discriminant functions for the normal density
- Error Bounds

Materials: To read relevant portions from

- Duda and Hart, Chapter 2
- Mitchell, Chapter 6

Bayes Rule

- Consider a two category classification into two classes c_1 and c_2.
- Let $P(c_i)$ and $p(\boldsymbol{x}|c_i)$ denote the prior probabilities and the class conditional probabilities respectively.
- Bayes Rule

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

Bayes Rule (contd.)

- Naturally if we have an example $\boldsymbol{x}$ such that

$$P(c_1|\boldsymbol{x}) > P(c_2|\boldsymbol{x}),$$

we would be inclined to assign class c_1 to $\boldsymbol{x}$.

- Probability of error:

$$\begin{aligned} P(error|\boldsymbol{x}) &= P(c_1|\boldsymbol{x}) \text{ if we decide } c_2 \\ & P(c_2|\boldsymbol{x}) \text{ if we decide } c_1 \end{aligned}$$

- Clearly by deciding c_1 if $P(c_1|\boldsymbol{x}) > P(c_2|\boldsymbol{x})$ and c_2 otherwise, we can minimize the probability of error.

Multicategory Case

- Suppose there are k classes $c_1, c_2, \ldots c_k$ are present thus for each of the k classes one can calculate $P_i = P(c_i|\boldsymbol{x})$
- Decide c_l as the class of $\boldsymbol{x}$ if

$$p_l = \max p_i$$

- Out of many ways to represent a classifier, one possible way is through **discriminant functions**.
- Thus for the k classes we can calculate k discriminant functions $g_i(\boldsymbol{x})$, $i = 1, 2, \ldots, k$.
- Decide the class label c_l to $\boldsymbol{x}$ if

$$g_l(\boldsymbol{x}) > g_i(\boldsymbol{x}), \forall i \neq l$$

Discriminant Functions

- Discriminant functions are not unique
- We can generally replace a discriminant function $g(\boldsymbol{x})$ by $f(g(\boldsymbol{x}))$, where $f(.)$ is a monotone increasing function.
- Thus for the Bayesian minimum error rate classification we can have the following equivalent discriminant functions:

$$\begin{aligned} g_i(\boldsymbol{x}) &= \frac{p(\boldsymbol{x}|c_i)P(c_i)}{p(\boldsymbol{x})} \\ &= p(\boldsymbol{x}|c_i)P(c_i) \\ &= \ln(p(\boldsymbol{x}|c_i)) + \ln P(c_i) \end{aligned}$$

Discriminant Funcs. (Contd.)

- Note, in the two category case, it is conventional to have a single discriminant function as we had in the case of logistic regression.
- Instead of using two different discriminant functions $g_1(\boldsymbol{x})$ and $g_2(\boldsymbol{x})$, it is more common to define a single function

$$g(\boldsymbol{x}) = g_1(\boldsymbol{x}) - g_2(\boldsymbol{x})$$

- Using this discriminant function we decide class c_1 if

$$g(\boldsymbol{x}) > 0,$$

and decide class c_2 otherwise.

The Normal Density

- Univariate normal density

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

- The expected value of x for this density is

$$\mu = \mathcal{E}[x] = \int_{-\infty}^{\infty} xp(x)dx.$$

- The expected squared deviation or variance is

$$\sigma^2 = \mathcal{E}[(x-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 p(x)dx.$$

The Normal Density (contd.)

- Multivariate normal density

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

  where,
  - $\boldsymbol{x} \in \mathcal{R}^d$
  - $\boldsymbol{\mu} \in \mathcal{R}^d$ is the *mean vector*.
  - Σ is the $d \times d$ *covariance matrix*.
- The above equation is often abbreviated as

$$p(\boldsymbol{x}) \sim N(\boldsymbol{\mu}, \Sigma)$$

The Normal Density (contd.)

- Formally we have

$$\boldsymbol{\mu} = \mathcal{E}[\boldsymbol{x}] = \int \boldsymbol{x} p(\boldsymbol{x}) d\boldsymbol{x}$$

and

$$\Sigma = \mathcal{E}[(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T] = \int (\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T p(\boldsymbol{x}) d\boldsymbol{x}$$

- **Note:** The expected value of a matrix or vector is found by taking the expected values of its components.

The Normal Density (contd.)

- Properties of Σ
  - The covariance matrix Σ is always symmetric and positive semidefinite. For our cases we shall consider Σ to be positive definite and thus $|\Sigma| > 0$.
  - The diagonal entries σ_{ii} are the variances of the respective x_i-s and the off-diagonal elements are the co-variances of x_i and x_j.
  - If x_i and x_j are statistically independent, then $\sigma_{ij} = 0$.
  - If all the off diagonal entries of Σ are 0 then $p(\boldsymbol{x})$ reduces to the product of the univariate densities of the components of $\boldsymbol{x}$.

DFs for Normal Density

- We saw that minimum error rate classification can be achieved by use of discriminant functions of the form

$$g_i(\boldsymbol{x}) = \ln p(\boldsymbol{x}|c_i) + \ln P(c_i)$$

- In case $p(\boldsymbol{x}|c_i) \sim N(\boldsymbol{\mu}_i, \Sigma_i)$, the discriminant functions can be easily evaluated. The form of the discriminant function then becomes:

$$g_i(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)^T \Sigma_i^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_i) - \frac{d}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma_i| + \ln P(c_i)$$

DFs for Normal Density(cont.)

- We shall investigate some special cases
  - Case 1: $\Sigma_i = \sigma^2 I$ yields linear boundary
  - Case 2: $\Sigma_i = \Sigma$ yields linear boundary
  - Case 3: Arbitrary Σ_i yields hyperquadrics

Error Bounds

- Consider a two class classification scenario.
- Suppose a classifier has partitioned the feature space into two regions R_1 and R_2 corresponding to the two classes c_1 and c_2.
- In this scenario, the probability of error would be

$$
\begin{aligned}
& P(error) \\
= \ & P(\boldsymbol{x} \in R_2, c_1) + P(\boldsymbol{x} \in R_1, c_2) \\
= \ & P(\boldsymbol{x} \in R_2 | c_1) P(c_1) + P(\boldsymbol{x} \in R_1 | c_2) P(c_2) \\
= \ & \int_{R_2} p(\boldsymbol{x}|c_1) P(c_1) d\boldsymbol{x} + \int_{R_1} p(\boldsymbol{x}|c_2) P(c_2) d\boldsymbol{x}
\end{aligned}
$$

Error Bounds (contd.)

- The probability of error can be written as

$$\begin{aligned} P(error) &= \int P(error, \boldsymbol{x}) d\boldsymbol{x} \\ &= \int P(error|\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} \end{aligned}$$

- Now, $P(error|\boldsymbol{x}) = \min[P(c_1|\boldsymbol{x}), P(c_2|\boldsymbol{x})]$
- Also we have

$$\min[a, b] \leq a^{\beta} b^{1-\beta}, \quad \text{for } a, b \geq 0 \ \text{ and } 0 \leq \beta \leq 1$$

Error Bounds (contd.)

- Combining we have

$$P(error) \leq P^{\beta}(c_1)P^{1-\beta}(c_2)\int p^{\beta}(\boldsymbol{x}|c_1)p^{1-\beta}(\boldsymbol{x}|c_2$$

- If the conditional probabilities are normal, then we can compute the integral analytically which gives

$$\int p^{\beta}(\boldsymbol{x}|c_1)p^{1-\beta}(\boldsymbol{x}|c_2)d\boldsymbol{x} = e^{-k(\beta)}$$

Error Bounds (contd.)

$$k(\beta) = \frac{\beta(1-\beta)}{2}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T[\beta\Sigma_1 + (1-\beta)\Sigma_2]^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) + \frac{1}{2}\ln\frac{|\beta\Sigma_1 + (1-\beta)\Sigma_2|}{|\Sigma_1|^{\beta}|\Sigma_2|^{(1-\beta)}}.$$

- The minimum value of $e^{-k(\beta)}$ gives the **Chernoff Bound** on the error probability.
- A less sharper bound called the **Bhattacharyya Bound** is obtained by substituting $\beta = 0.5$

Issues to be addressed in next class

- In real life problems, we generally do not have access to the probability values.
- How to estimate the probabilities from data?
- Refer Chapter 3 of Duda and Hart and Chapter 6 of Mitchell
- We shall discuss such techniques in the next class