Linear Discrimination

Ziping Zhao

School of Information Science and Technology
ShanghaiTech University, Shanghai, China

CS182: Introduction to Machine Learning (Fall 2022)
http://cs182.sist.shanghaitech.edu.cn

Ch. 10 of I2ML (Sec. 10.8 & Sec. 10.9 excluded)

Outline

Outline

Likelihood-Based vs. Discriminant-Based Classification – I

- Classification based on a set of discriminant functions $g_i(x)$, $i = 1, \ldots, K$:

$$\text{Choose } C_i \text{ if } g_i(\mathbf{x}) = \max_k g_k(\mathbf{x})$$

- Likelihood-based classification:
  - Assume a parametric, semiparametric, or nonparametric model for the class-conditional probability densities $p(\mathbf{x} \mid C_i)$.
  - Estimate the prior probabilities $P(C_i)$ and the class likelihoods $p(\mathbf{x} \mid C_i)$ from data.
  - Apply Bayes' rule to compute the posterior probabilities $P(C_i \mid \mathbf{x})$.
  - Perform optimal classification based on $P(C_i \mid \mathbf{x})$, or equivalently based on discriminant functions $g_i(\mathbf{x})$ such as $g_i(x) = \log P(C_i \mid \mathbf{x})$.
- Discriminant-based classification:
  - Assume a model directly for the discriminant functions, bypassing the estimation of $p(\mathbf{x} \mid C_i)$ or $P(C_i \mid \mathbf{x})$ from data.
  - Perform optimal classification based on the discriminant functions $g_i(\mathbf{x})$.

Likelihood-Based vs. Discriminant-Based Classification – II

- Main difference:
  - the likelihood-based approach makes an assumption on the form of the densities (e.g., whether they are Gaussian, or whether the inputs are correlated, etc.)
  - the discriminant-based approach makes an assumption on the form of the discriminants (e.g., whether they are linear)

Discriminant Functions

- Define a model for the discriminant function of class C_i:

$$g_i(\mathbf{x} \mid \Phi_i)$$

  which are explicitly parameterized with a set of model parameters Φ_i.
  - In discriminant-based approach, we make an assumption on the form of the boundaries separating classes.
- Learning is the optimization of Φ_i to maximize the quality of the separation, that is, the classification accuracy on a given labeled training set.
- Unlike the likelihood-based approach which performs density estimation separately for each class, the discriminant-based approach typically estimates Φ_i for all classes simultaneously to find the decision boundaries between classes.
- Estimating the class boundaries (i.e., the discriminants) is usually easier than estimating the class densities.
  - E.g., this is true when the discriminant can be approximated by a simple function.

Linear Discriminant Functions

- Linear discriminant function:

$$g_i(\mathbf{x} \mid \mathbf{w}_i, w_{i0}) = \mathbf{w}_i^T \mathbf{x} + w_{i0} = \sum_{j=1}^{d} w_{ij} x_j + w_{i0}$$

  which is linear in $\mathbf{x}$.
- Advantages:
  - Simplicity: $O(d)$ time and space complexity.
  - Understandability: final output is a weighted sum of attributes; magnitude and sign of weights have clear physical meaning.
  - Accuracy: model is quite accurate if some assumptions are satisfied, e.g., Gaussian densities for classes with shared covariance matrix.
- We should always use the linear discriminant before trying a more complicated model to make sure that the additional complexity is justified.

Generalizing the Linear Models

- When a linear model is not flexible enough, we can use the quadratic discriminant function
$$g_i(\mathbf{x} \mid \mathbf{W}_i, \mathbf{w}_i, w_{i0}) = \mathbf{x}^T \mathbf{W}_i \mathbf{x} + \mathbf{w}_i^T \mathbf{x} + w_{i0}$$
- $O(d^2)$ time and space complexity
- An equivalent way is to preprocess the input by adding higher-order terms (or called product terms).
  - Example: with two inputs x_1 and x_2, we define new variables
$$z_1 = x_1,\; z_2 = x_2\; z_3 = x_1^2,\; z_4 = x_2^2,\; z_5 = x_1 x_2$$
    and take $\mathbf{z} = (z_1, z_2, z_3, z_4, z_5)^T$ as the new input. The linear function defined in the new $\mathbf{z}$-space corresponds to a nonlinear function in the original $\mathbf{x}$-space.
- Compared with defining a nonlinear function (discriminant or regression) in the original input space, defining a linear function in a nonlinearly transformed new space (called a generalized linear model) does not increase the number of parameters that need to be estimated significantly.

Basis Functions

- More generally, the inputs $\mathbf{x}$ are (nonlinearly) transformed into basis functions $\phi_{ij}(\mathbf{x})$ which are linearly combined to define the discriminant functions:

$$g_i(\mathbf{x}) = \mathbf{w}^T \boldsymbol{\phi}_i(\mathbf{x}) = \sum_{j=1}^{k} w_j \phi_{ij}(\mathbf{x})$$

- Higher-order terms mentioned before are only one set of basis functions.
- Other examples of basis functions:
  - $\sin(x_1)$
  - $\exp(-(x_1 - m)^2/c)$
  - $\exp(-\|\mathbf{x} - \mathbf{m}\|^2/c)$
  - $\log(x_1)$
  - $\mathbf{1}(x_1 > c)$
  - $\mathbf{1}(ax_1 + bx_2 > c)$
  - ...

Outline

Geometric View: Two Classes

- Discriminant function:

$$
\begin{aligned}
g(\mathbf{x}) &= g_1(\mathbf{x}) - g_2(\mathbf{x}) \\
&= (\mathbf{w}_1^T \mathbf{x} + w_{10}) - (\mathbf{w}_2^T \mathbf{x} + w_{20}) \\
&= (\mathbf{w}_1 - \mathbf{w}_2)^T \mathbf{x} + (w_{10} - w_{20}) \\
&= \mathbf{w}^T \mathbf{x} + w_0
\end{aligned}
$$

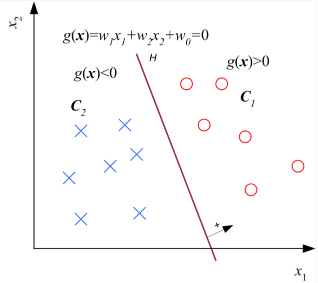
  where $\mathbf{w}$ is the weight vector and w_0 is the threshold.

- Optimal decision rule:

$$
\text{Choose } \begin{cases} C_1 & \text{if } g(\mathbf{x}) > 0 \\ C_2 & \text{otherwise} \end{cases}
$$

Hyperplane

- The discriminant function defines a hyperplane H, i.e., $g(\mathbf{x}) = 0$, that divides the input space into 2 half-spaces:
  - Decision region $\mathcal{R}_1$ for C_1 ($g(\mathbf{x}) > 0$, i.e., positive side of the hyperplane)
  - Decision region $\mathcal{R}_2$ for C_2 ($g(\mathbf{x}) < 0$, i.e., negative side of the hyperplane)



- When $\mathbf{x} = \mathbf{0}$ (i.e., the origin), $g(\mathbf{x}) = w_0$. If $w_0 > 0$, the origin is on the positive side, and if $w_0 < 0$, the origin is on the negative side, and if $w_0 = 0$, the hyperplane passes through the origin.

Geometric Interpretation – I

- Let $\mathbf{x}_1$ and $\mathbf{x}_2$ be two points on the hyperplane, i.e., $g(\mathbf{x}_1) = g(\mathbf{x}_2) = 0$. So

$$\mathbf{w}^T\mathbf{x}_1 + w_0 = \mathbf{w}^T\mathbf{x}_2 + w_0$$
$$\mathbf{w}^T(\mathbf{x}_1 - \mathbf{x}_2) = 0$$

  showing that $\mathbf{w}$ is normal (or orthogonal) to any vector $\mathbf{x}_1 - \mathbf{x}_2$ lying on the hyperplane.

- Let us express any point $\mathbf{x}$ as

$$\mathbf{x} = \mathbf{x}_p + r\frac{\mathbf{w}}{\|\mathbf{w}\|}$$

  where

  $\mathbf{x}_p$: normal projection of $\mathbf{x}$ onto the hyperplane

  r : distance from $\mathbf{x}$ to the hyperplane ($r > / < 0$: $\mathbf{x}$ is on the positive/negative side)

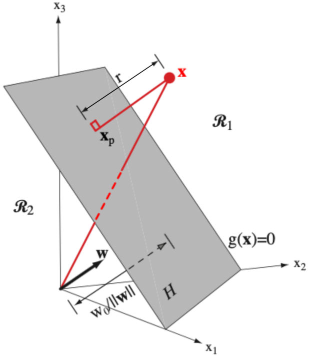
Geometric Interpretation – II

- Calculation of r (note $g(\mathbf{x}_p) = 0$):

$$g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0 = \mathbf{w}^T\mathbf{x}_p + r\frac{\mathbf{w}^T\mathbf{w}}{\|\mathbf{w}\|} + w_0 = g(\mathbf{x}_p) + r\|\mathbf{w}\| = r\|\mathbf{w}\|$$

So we have

$$r = \frac{g(\mathbf{x})}{\|\mathbf{w}\|} \qquad (\text{sign of } r = \text{sign of } g(\mathbf{x}))$$



Geometric Interpretation – III

- When $\mathbf{x} = \mathbf{0}$, the distance from origin to hyperplane is $\frac{g(\mathbf{0})}{\|\mathbf{w}\|} = \frac{w_0}{\|\mathbf{w}\|}$.
  - Alternative proof: if $\mathbf{x}$ is a point on the hyperplane, then $g(\mathbf{x}) = 0$. So

$$
\begin{aligned}
\mathbf{w}^T\mathbf{x} + w_0 &= 0 \\
\left(\frac{\mathbf{w}}{\|\mathbf{w}\|}\right)^T \mathbf{x} + \frac{w_0}{\|\mathbf{w}\|} &= 0 \\
\left|\left(\frac{\mathbf{w}}{\|\mathbf{w}\|}\right)^T \mathbf{x}\right| &= \frac{w_0}{\|\mathbf{w}\|}
\end{aligned}
$$

- The orientation of the hyperplane is determined by $\mathbf{w}$ and its distance from the origin is determined by w_0 and $\mathbf{w}$.

Geometric View: Multiple Classes

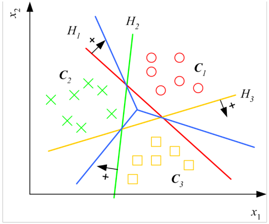
- K discriminant functions:

$$g_i(\mathbf{x} \mid \mathbf{w}_i, w_{i0}) = \mathbf{w}_i^T \mathbf{x} + w_{i0}$$

- Linearly separable classes:

$$g_i(\mathbf{x} \mid \mathbf{w}_i, w_{i0}) = \begin{cases} > 0 & \text{if } \mathbf{x} \in C_i \\ \leq 0 & \text{otherwise} \end{cases}$$

For each class C_i, there exists a hyperplane H_i such that all $\mathbf{x} \in C_i$ lie on the positive side and all other $\mathbf{x} \in C_j$, $j \neq i$ lie on the negative side.



Linear Classifier

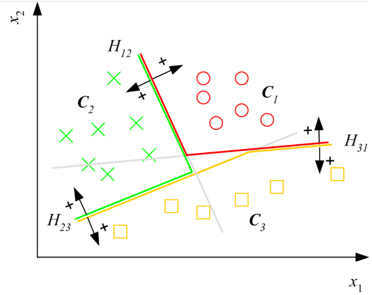
- During testing, given $\mathbf{x}$, ideally, we should have only one $g_j(\mathbf{x})$, $j = 1, \ldots, K$ greater than 0.
- However, it is possible for multiple or no $g_i(\mathbf{x})$ to be > 0. These may be taken as reject cases, but the usual approach is to assign $\mathbf{x}$ to the class having the highest discriminant.
- Decision rule for any test case $\mathbf{x}$:

$$\text{Choose } C_i \text{ if } g_i(\mathbf{x}) = \max_{j=1}^{K} g_j(\mathbf{x})$$

- Geometrically, a linear classifier partitions the feature space into K convex decision regions $\mathcal{R}_i$.

Pairwise Separation – I

- If the classes are not linearly separable, one approach is to divide it into a set of linear problems and linear discriminants can be used to separate the classes.
- One possibility is to perform pairwise separation of classes by considering one pair of distinct classes at a time.
- $K(K-1)/2$ linear discriminants are used.
- It is easier for the classes to be pairwise linearly separable than linearly separable.



Pairwise Separation – II

- Discriminant function for classes i and j ($i, j = 1, \ldots, K$ and $j \neq i$):

$$g_{ij}(\mathbf{x} \mid \mathbf{w}_{ij}, \mathbf{w}_{ij0}) = \mathbf{w}_{ij}^T \mathbf{x} + w_{ij0} = \begin{cases} > 0 & \text{if } \mathbf{x} \in C_i \\ \leq 0 & \text{if } \mathbf{x} \in C_j \\ \text{don't care} & \text{if } \mathbf{x} \in C_k, k \neq i, k \neq j \end{cases}$$

  - if $\mathbf{x}^t \in C_k$ where $k \neq i$, $k \neq j$, then $\mathbf{x}^t$ is not used during training of $g_{ij}(\mathbf{x})$.
- Decision rule for any test case $\mathbf{x}$:

$$\text{Choose } C_i \text{ if } \forall j \neq i, g_{ij}(\mathbf{x}) > 0$$

- Sometimes we may not be able to find such a class C_i. If we do not want to reject such cases, a relaxed decision rule can be defined based on a new set of discriminant functions:

$$g_i(\mathbf{x}) = \sum_{j \neq i} g_{ij}(\mathbf{x})$$

Outline

Linear Parametric Classification Revisited - I

- Recall that if the class-conditional densities $p(\mathbf{x} \mid C_i)$ are Gaussian sharing a common covariance matrix $\mathbf{\Sigma}$, the discriminant functions are linear:

$$g_i(\mathbf{x}) = \mathbf{w}_i^T \mathbf{x} + w_{i0}$$

where

$$\mathbf{w}_i = \mathbf{\Sigma}^{-1} \boldsymbol{\mu}_i$$

$$w_{i0} = -\frac{1}{2} \boldsymbol{\mu}_i^T \mathbf{\Sigma}^{-1} \boldsymbol{\mu}_i + \log P(C_i)$$

- Given a sample $\mathcal{X}$, we find the ML estimates for $\boldsymbol{\mu}_i$ and $\mathbf{\Sigma}$, denoted by $\mathbf{m}_i$ and $\mathbf{S}$, and plug them into the discriminant functions.

Linear Parametric Classification Revisited - II

- For two class classification, we have the discriminant function:

$$
\begin{aligned}
g(\mathbf{x}) &= g_1(\mathbf{x}) - g_2(\mathbf{x}) \\
&= (\mathbf{w}_1^T\mathbf{x} + w_{10}) - (\mathbf{w}_2^T\mathbf{x} + w_{20}) \\
&= (\mathbf{w}_1 - \mathbf{w}_2)^T\mathbf{x} + (w_{10} - w_{20}) \\
&= \mathbf{w}^T\mathbf{x} + w_0
\end{aligned}
$$

where

$$
\mathbf{w} = \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), \quad w_0 = -\frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)^T\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \log\frac{P(C_1)}{P(C_2)}
$$

- Optimal decision rule:

$$
\text{Choose} \begin{cases} C_1 & \text{if } g(\mathbf{x}) > 0 \\ C_2 & \text{otherwise} \end{cases}
$$

Two-Class Example

- Let

$$P(C_1 \mid \mathbf{x}) = y \qquad P(C_2 \mid \mathbf{x}) = 1 - y$$

- Classification rule:

$$\text{Choose } \begin{cases} C_1 & \text{if } y > 0.5 \\ C_2 & \text{otherwise} \end{cases}$$

- Equivalent tests for classification rule:

$$\frac{y}{1-y} > 1 \qquad \textit{or} \qquad \log \frac{y}{1-y} > 0$$

  where
  - $\frac{y}{1-y}$ is called the odds (or odds ratio) of y
  - $\log \frac{y}{1-y}$ is called the log-odds of y or logit (logistic unit) transformation/function of y, written as $\mathrm{logit}(y)$.
- The $\mathrm{logit}(\cdot)$ is a type of function that maps probability values from $(0, 1)$ to real numbers in $(-\infty, +\infty)$.

Logit Function

- In the case of two normal classes sharing a common covariance matrix, the logit function:

$$\begin{aligned}
\operatorname{logit}(P(C_1 \mid \mathbf{x})) &= \log \frac{P(C_1 \mid \mathbf{x})}{1 - P(C_1 \mid \mathbf{x})} = \log \frac{P(C_1 \mid \mathbf{x})}{P(C_2 \mid \mathbf{x})} \\
&= \log \frac{p(\mathbf{x} \mid C_1)}{p(\mathbf{x} \mid C_2)} + \log \frac{P(C_1)}{P(C_2)} \\
&= \log \frac{(2\pi)^{-\frac{d}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_1)]}{(2\pi)^{-\frac{d}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_2)]} + \log \frac{P(C_1)}{P(C_2)} \\
&= \mathbf{w}^T \mathbf{x} + w_0
\end{aligned}$$

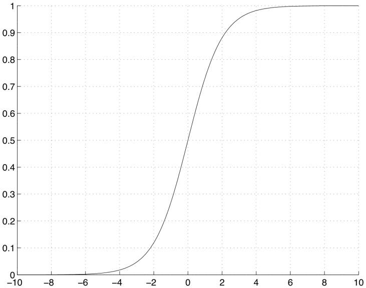
where

$$\mathbf{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), \quad w_0 = -\frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \log \frac{P(C_1)}{P(C_2)}$$

Sigmoid Function – I

- Sigmoid function or logistic function (inverse function of logit):

$$\text{sigmoid}(a) = \frac{1}{1 + \exp(-a)}$$



Sigmoid Function – II

- Then,

$$P(C_1 \mid \mathbf{x}) = \text{sigmoid}(\mathbf{w}^T\mathbf{x} + w_0) = \frac{1}{1 + \exp[-(\mathbf{w}^T\mathbf{x} + w_0)]}$$

  which directly computes the posterior class probability $P(C_1 \mid \mathbf{x})$.
- Training:
  - Estimate $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_1$, and $\boldsymbol{\Sigma}$ from data and plug the estimates into the discriminant functions.
- Testing:
  - Calculate $g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0$ and choose C_1 if $g(x) > 0$, or
  - Calculate $y = \text{sigmoid}(\mathbf{w}^T\mathbf{x} + w_0)$ and choose C_1 if $y > 0.5$ (since y can be interpreted as a posterior probability and $\text{sigmoid}(0) = 0.5$).