Linear classifier: a machine learning algorithm that makes classification decisions based on a linear combination of characteristics of a given data point.

- If the input feature vector is a real vector $\vec{x}$, the output is:

$$y = f(\vec{w}^T \vec{x})$$

where $\vec{w}$ is a real vector of weights and f is a function that converts $\vec{w}^T\vec{x}$ into the desired output. (i.e. class of data)

- The weight vector is learned from training examples.
- The function f is often a threshold function that maps values of $\vec{w}^T\vec{x}$ above a certain threshold to the first class and other values to the second class.

$$f(\vec{w}^T\vec{x}) = \begin{cases} 1 & \text{if } \vec{w}^T\vec{x} > T \\ 0 & \text{otherwise} \end{cases} \quad (T \text{ is the threshold})$$

★ f could also be a function that outputs probability that an item belongs to a certain class.

Examples:

Generative Models: models conditional density function $P(\vec{x} \mid C)$. It models how a particular class would generate input data. It tries to learn the decision boundary for the model.

$$P(C_k \mid \vec{x}) = \frac{P(\vec{x} \mid C_k) P(C_k)}{\sum_{j=1}^{n} P(\vec{x} \mid C_j) P(C_j)}$$

Linear Discriminant Analysis: Assumes Gaussian conditional density models.

Naive Bayes classifier: family of sample "probabilistic" classifiers based on applying Baye's theorem with strong independence assumptions b/w families.

Discriminitive Models: Directly calculates $P(C \mid \vec{x})$ to learn a direct map from input x to label y. These models learn what the features in the input are most useful to distinguish between the various possible classes.

Logistic Regression: ML estimate of $\vec{w}$ assuming training set was generated by a binomial model that depends on the output of the classifier.

Fisher's LDA: maximizes ratio of between-class scatter to within-class scatter. Large mean separation between classes, Small projected with-in class variance

Support Vector machine: maximizes the margin between the decision hyperplane. and training examples.

Fisher's LDA:

Goal: Project data into 1 dimension to achieve a large class separation.

maximize $$\frac{\vec{w}^T S_B \vec{w}}{\vec{w}^T S_W \vec{w}}$$ (see class notes for detail)

where

$$S_B = (\vec{m}_2 - \vec{m}_1)(\vec{m}_2 - \vec{m}_1)^T$$ Between class Covariance.

$$S_W = \sum_{\vec{x}_i \in C_1} (\vec{x}_i - \vec{m}_1)(\vec{x}_i - \vec{m}_1)^T + \sum_{\vec{x}_i \in C_2} (\vec{x}_i - \vec{m}_2)(\vec{x}_i - \vec{m}_2)^T$$ Total with-in class Covariance Matrix.

Solution to $\vec{w}$:

$$\vec{w} = S_W^{-1}(\vec{m}_2 - \vec{m}_1)$$

bias term: no general rule.

often use $\vec{b} = \frac{1}{2}(\vec{w}^T\vec{m}_1 + \vec{w}^T\vec{m}_2)$

Example problem:

Suppose we have 4 data points, 2 points in class 1: $x_1 = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$, $x_2 = \begin{bmatrix} 5 \\ 2 \end{bmatrix}$

2 points in class 2: $x_3 = \begin{bmatrix} 1 \\ 4 \end{bmatrix}$ $x_4 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}$

① Compute the mean vectors for the 2 classes.

$$\vec{m}_1 = \frac{1}{2}\begin{bmatrix} 3+5 \\ 2+2 \end{bmatrix} = \begin{bmatrix} 4 \\ 2 \end{bmatrix}$$

$$\vec{m}_2 = \frac{1}{2}\begin{bmatrix} 1+3 \\ 4+6 \end{bmatrix} = \begin{bmatrix} 2 \\ 5 \end{bmatrix}$$

② Compute the between-class covariance Matrix.

covariance Matrix of class 1:

$$\sum_{\vec{x}_i \in C_1} (\vec{x}_i - \vec{m}_1)(\vec{x}_i - \vec{m}_1)^T$$

$$= \left[\left(\begin{bmatrix} 3 \\ 2 \end{bmatrix} - \begin{bmatrix} 4 \\ 2 \end{bmatrix}\right)\left(\begin{bmatrix} 3 \\ 2 \end{bmatrix} - \begin{bmatrix} 4 \\ 2 \end{bmatrix}\right)^T + \left(\begin{bmatrix} 5 \\ 2 \end{bmatrix} - \begin{bmatrix} 4 \\ 2 \end{bmatrix}\right)\left(\begin{bmatrix} 5 \\ 2 \end{bmatrix} - \begin{bmatrix} 4 \\ 2 \end{bmatrix}\right)^T\right]$$

$$= \left[\left(\begin{bmatrix} -1 \\ 0 \end{bmatrix} \cdot [-1, 0]\right) + \left(\begin{bmatrix} 1 \\ 0 \end{bmatrix} [1, 0]\right)\right]$$

$$= \left(\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}\right)$$

$$= \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}$$

Covariance Matrix of class 2:

$$\sum_{\vec{x}_i \in C_2} (\vec{x}_i - \vec{m}_2)(\vec{x}_i - \vec{m}_2)^T$$

$$= \left[\left(\begin{bmatrix} 1 \\ 4 \end{bmatrix} - \begin{bmatrix} 2 \\ 5 \end{bmatrix}\right)\left(\begin{bmatrix} 1 \\ 4 \end{bmatrix} - \begin{bmatrix} 2 \\ 5 \end{bmatrix}\right)^T + \left(\begin{bmatrix} 3 \\ 6 \end{bmatrix} - \begin{bmatrix} 2 \\ 5 \end{bmatrix}\right)\left(\begin{bmatrix} 3 \\ 6 \end{bmatrix} - \begin{bmatrix} 2 \\ 5 \end{bmatrix}\right)^T\right]$$

$$= \left[\left(\begin{bmatrix} -1 \\ -1 \end{bmatrix}[-1, -1]\right) + \left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}[1, 1]\right)\right]$$

$$= \left(\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} + \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \right)$$

$$= \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}$$

$$S_w = \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix} = \begin{bmatrix} 4 & 2 \\ 2 & 2 \end{bmatrix}$$

③ compute $\vec{w}$

$$\vec{w} = S_w^{-1} (\vec{m_2} - \vec{m_1})$$

inverse of 2x2 matrix: $\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad-bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$

$$S_w^{-1} = \frac{1}{8-4} \begin{bmatrix} 2 & -2 \\ -2 & 4 \end{bmatrix}$$

$$= \frac{1}{4} \begin{bmatrix} 2 & -2 \\ -2 & 4 \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{bmatrix}$$

$$\vec{w} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{bmatrix} \left(\begin{bmatrix} 2 \\ 5 \end{bmatrix} - \begin{bmatrix} 4 \\ 2 \end{bmatrix} \right)$$

$$= \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{bmatrix} \cdot \begin{bmatrix} -2 \\ 3 \end{bmatrix}$$

$$= \begin{bmatrix} -1-1.5 \\ -1+3 \end{bmatrix}$$

$$= \begin{bmatrix} -2.5 \\ 2 \end{bmatrix}$$

Linear Discriminant Analysis: extension of fisher's Discriminant Analysis to multiple classes.

Assumptions:

① Gaussian distribution for $P(\vec{x}|C_k)$

② Covariance matrix is the same for all classes.

get parameters by doing maximum likelihood estimation

$$\vec{w} = \underset{\text{shared covariance matrix.}}{\Sigma^{-1}} (\vec{\mu}_1 - \vec{\mu}_2) \qquad b = -\frac{1}{2}\vec{\mu}_1^T \Sigma^{-1} \vec{\mu}_1 + \frac{1}{2}\vec{\mu}_2^T \Sigma^{-1} \vec{\mu}_2$$

(This is for 2-class case)

where $\Sigma = \frac{n_1}{n} S_1 + \frac{n-n_1}{n} S_2$.

and $S_1 = \frac{1}{n_1} \sum_{\vec{x}_i \in C_1} (\vec{x}_i - \vec{\mu}_1)(\vec{x}_i - \vec{\mu}_1)^T$

$S_2 = \frac{1}{n-n_1} \sum_{\vec{x}_i \in C_2} (\vec{x}_i - \vec{\mu}_2)(\vec{x}_i - \vec{\mu}_2)^T$

example problem:

Suppose we have 4 training points. 2 points in class 1:

$$x_1 = \begin{bmatrix} 2 \\ 4 \end{bmatrix} \qquad x_2 = \begin{bmatrix} 4 \\ 6 \end{bmatrix}$$

2 points in class 2:

$$x_3 = \begin{bmatrix} 1 \\ 6 \end{bmatrix} \qquad x_4 = \begin{bmatrix} 3 \\ 8 \end{bmatrix}$$

① Compute the mean vectors μ_1 and μ_2 for the 2 classes.

$$\mu_1 = \frac{1}{2}\begin{bmatrix}(2+4)\\(4+6)\end{bmatrix} = \begin{bmatrix}3\\5\end{bmatrix}$$

$$\mu_2 = \frac{1}{2}\begin{bmatrix}(1+3)\\(6+8)\end{bmatrix} = \begin{bmatrix}2\\7\end{bmatrix}$$

② Compute the shared covariance matrix Σ

$$S_1 = \frac{1}{n_1}\sum_{\vec{x}_i \in C_1} (\vec{x}_i - \vec{\mu}_1)(\vec{x}_i - \vec{\mu}_1)^T.$$

$$= \frac{1}{2}\left[\left(\begin{bmatrix}2\\4\end{bmatrix} - \begin{bmatrix}3\\5\end{bmatrix}\right)\left(\begin{bmatrix}2\\4\end{bmatrix} - \begin{bmatrix}3\\5\end{bmatrix}\right)^T + \left(\begin{bmatrix}4\\6\end{bmatrix} - \begin{bmatrix}3\\5\end{bmatrix}\right)\left(\begin{bmatrix}4\\6\end{bmatrix} - \begin{bmatrix}3\\5\end{bmatrix}\right)^T\right]$$

$$= \frac{1}{2}\left[\left(\begin{bmatrix}-1\\-1\end{bmatrix}[-1, -1]\right) + \left(\begin{bmatrix}1\\1\end{bmatrix}[1, 1]\right)\right]$$

$$= \frac{1}{2}\left(\begin{bmatrix}1 & 1\\1 & 1\end{bmatrix} + \begin{bmatrix}1 & 1\\1 & 1\end{bmatrix}\right)$$

$$= \begin{bmatrix}1 & 1\\1 & 1\end{bmatrix}$$

$$S_2 = \frac{1}{n-n_1}\sum_{\vec{x}_i \in C_2} (\vec{x}_i - \vec{\mu}_2)(\vec{x}_i - \vec{\mu}_2)^T$$

$$= \frac{1}{2}\left[\left(\begin{bmatrix}1\\6\end{bmatrix} - \begin{bmatrix}2\\7\end{bmatrix}\right)\left(\begin{bmatrix}1\\6\end{bmatrix} - \begin{bmatrix}2\\7\end{bmatrix}\right)^T + \left(\begin{bmatrix}3\\8\end{bmatrix} - \begin{bmatrix}2\\7\end{bmatrix}\right)\left(\begin{bmatrix}3\\8\end{bmatrix} - \begin{bmatrix}2\\7\end{bmatrix}\right)^T\right]$$

$$= \frac{1}{2}\left[\left(\begin{bmatrix}-1\\-1\end{bmatrix}[-1, -1]\right) + \left(\begin{bmatrix}1\\1\end{bmatrix}[1, 1]\right)\right]$$

$$= \begin{bmatrix}1 & 1\\1 & 1\end{bmatrix}$$

$$\Sigma = \frac{1}{2} S_1 + \frac{1}{2} S_2$$
$$= \frac{1}{2}(S_1 + S_2)$$
$$= \frac{1}{2}\begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}$$
$$= \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$