Machine Learning I: Fractal 2

Rajendra Nagar

Assistant Professor
Department of Electircal Engineering
Indian Institute of Technology Jodhpur
http://home.iitj.ac.in/∼rn/

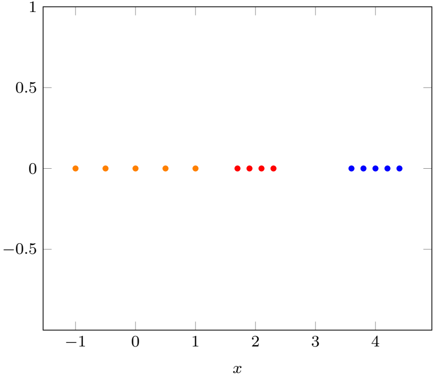
These slides are prepared from the following sources:

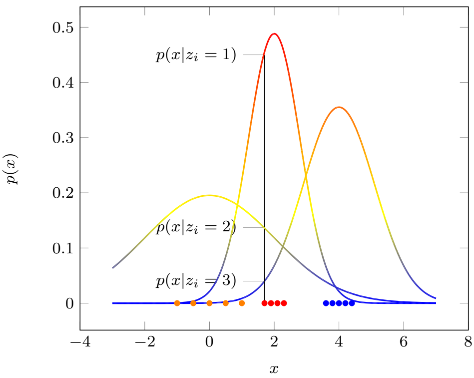
Bishop, C. M. (2006). Pattern recognition and machine learning. springer.
Shlens, J. (2014). A tutorial on independent component analysis. arXiv preprint arXiv:1404.2986.

Gaussian Mixture Models: Problem Formulation

Given a set $\mathcal{S} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ of n independent samples drawn from a mixture of k Gaussian distributions, find the following:

- Determine the probability of each point being sampled from a particular Gaussian distribution. (Cluster assignment probabilities, responsibilities)
- Estimate the parameters (mean vector and covariance matrix) for all k distributions. (Cluster representatives)





Expectation Maximization

1: **Input**: $\mathcal{S} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, where $p(\mathbf{x}) = \sum_{i \in [k]} \pi_i \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$.

2: **Maximize log-likelihood:** $\max_{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}} \sum_{j \in [n]} \log\left(\sum_{i \in [k]} \pi_i \mathcal{N}(\mathbf{x}_j|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\right)$

3: **Initialize:** $\boldsymbol{\mu}_t$, $\boldsymbol{\Sigma}_t$, and π_t, $\forall t \in [k]$.

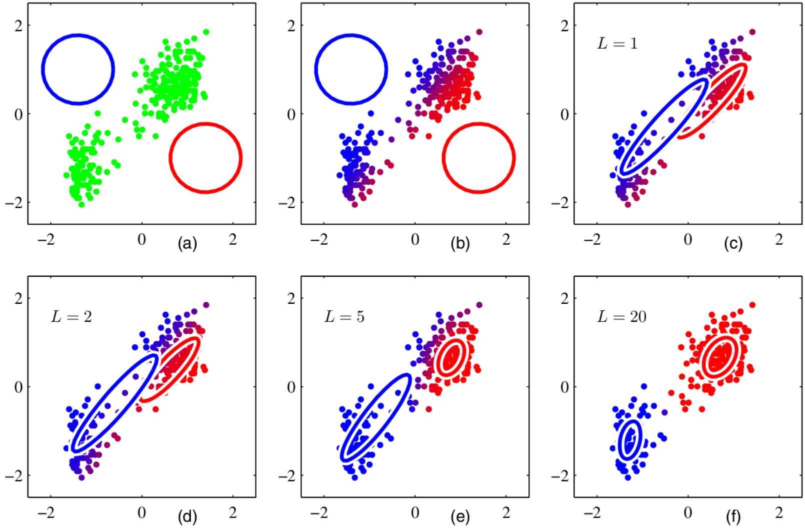
4: **E** step. Evaluate the responsibilities using the current parameter values.

$$\gamma(jt) \leftarrow \frac{\pi_t \mathcal{N}(\mathbf{x}_j|\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)}{\sum_{i \in [k]} \pi_i \mathcal{N}(\mathbf{x}_j|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}.$$

5: **M** step. Re-estimate the parameters using the current responsibilities.

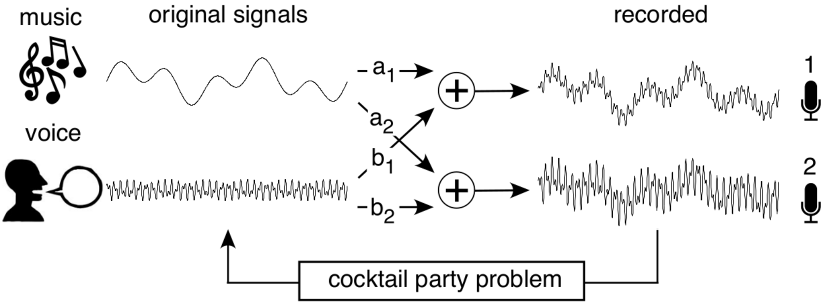
$$\begin{aligned} \boldsymbol{\mu}_t^{\text{new}} &\leftarrow \frac{1}{n_t} \sum_{j \in [n]} \gamma(jt)\mathbf{x}_j \\ \boldsymbol{\Sigma}_t^{\text{new}} &\leftarrow \frac{1}{n_t} \sum_{j \in [n]} \gamma(jt)(\mathbf{x}_j - \boldsymbol{\mu}_t^{\text{new}})(\mathbf{x}_j - \boldsymbol{\mu}_t^{\text{new}})^\top \\ \pi_t^{\text{new}} &\leftarrow \frac{n_t}{n}. \text{ Here, } n_t = \sum_{j \in [n]} \gamma(jt). \end{aligned}$$

6: Check the log-likelihood value $\log(\ell(\boldsymbol{\mu}_t^{\text{new}}, \boldsymbol{\Sigma}_t^{\text{new}}, \pi_t^{\text{new}}))$ for convergence.



2
0
−2
−2
0
(a)
2
2
0
−2
−2
0
(b)
2
$L = 1$
2
0
−2
−2
0
(c)
2
$L = 2$
2
0
−2
−2
0
(d)
2
$L = 5$
2
0
−2
−2
0
(e)
2
$L = 20$
2
0
−2
−2
0
(f)
2

Independent Component Analysis



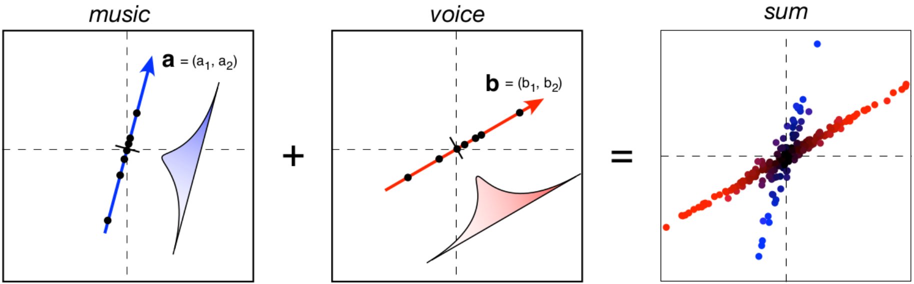
- Two sounds s_1, s_2 are generated by music and a voice and recorded simultaneously in two microphones. Sound adds linearly.
- Two microphones record a unique linear summation of the two sounds.
- The linear weights for each microphone (a_1 , b_1 and a_2 , b_2) reflect the proximity of each speaker to the respective microphones.
- The goal of the *cocktail party problem* is to recover the original sources (i.e. music and voice) using the microphone recording.

Independent Component Analysis

- Let $\mathbf{x} \in \mathbb{R}^{2\times n}$ be the observed data, $\mathbf{s} \in \mathbb{R}^{2\times n}$ be the original data, and let $\mathbf{A} = \begin{bmatrix} a_1 & b_1 \\ a_2 & b_2 \end{bmatrix}$ be the *mixing matrix* that is invertible and unknown. Then, we have that

$$\mathbf{x} = \mathbf{A}\mathbf{s} = \begin{bmatrix} a_1\mathbf{s}_1^\top \\ a_2\mathbf{s}_1^\top \end{bmatrix} + \begin{bmatrix} b_1\mathbf{s}_2^\top \\ b_2\mathbf{s}_2^\top \end{bmatrix}.$$

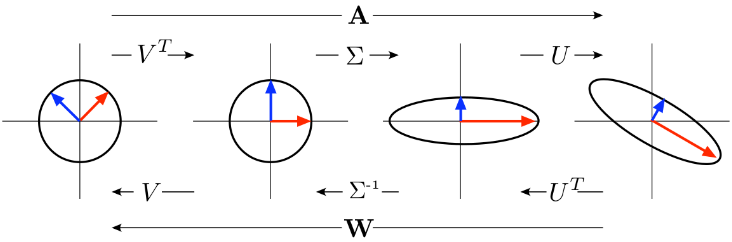
- Let $\mathbf{W} = \mathbf{A}^{-1}$ be the *unmixing matrix* that is $\hat{\mathbf{s}} = \mathbf{W}\mathbf{x}$.
- We only know $\mathbf{x}$. The matrix $\mathbf{W}$ and original data $\mathbf{s}$ are unknown (ill-posed problem).



Independent Component Analysis

- Rather than trying to solve for $\mathbf{s}$ and $\mathbf{A}$ simultaneously, we focus on finding $\mathbf{A}$ first.
- Rather than solving for $\mathbf{A}$, we solve for $\mathbf{A}$ by decomposing it into meaningful matrices.
- The singular value decomposition factorizes $\mathbf{A}$ into 3 geometrically meaningful matrices:

  $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top \Rightarrow \mathbf{W} = \mathbf{V}\boldsymbol{\Sigma}^{-1}\mathbf{U}^\top$
- The matrices $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{2\times 2}$ are orthonormal matrices (Rotation Matrices).
- The matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{2\times 2}$ is a diagonal matrix (Nonuniform Scaling).



Steps

- Use the covariance of the data $\mathbf{x}$ in order to calculate $\mathbf{U}$ and $\boldsymbol{\Sigma}$.
- Use statistical independence of $\mathbf{s}$ to solve for $\mathbf{V}$.

Independent Component Analysis

- Assume that the covariance matrix of $\mathbf{s}$ is the identity matrix (*whitened data*). That is:

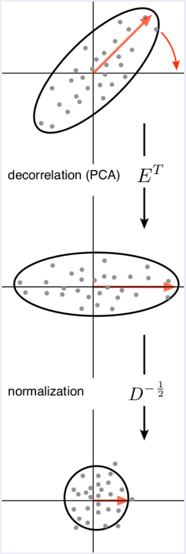
$$
\begin{aligned}
\mathbb{E}\left[\mathbf{s}\mathbf{s}^\top\right] &= \mathbf{I} \\
\mathbb{E}\left[\mathbf{x}\mathbf{x}^\top\right] &= \mathbb{E}\left[\mathbf{A}\mathbf{s}(\mathbf{A}\mathbf{s})^\top\right] \\
&= \mathbb{E}\left[\mathbf{A}\mathbf{s}\mathbf{s}^\top\mathbf{A}^\top\right] \\
&= \mathbf{A}\mathbb{E}\left[\mathbf{s}\mathbf{s}^\top\right]\mathbf{A}^\top \\
&= \mathbf{A}\mathbf{A}^\top \\
&= (\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top)(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top)^\top \\
&= \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top\mathbf{V}\mathbf{\Sigma}\mathbf{U}^\top \\
\mathbf{C}_\mathbf{x} &= \mathbf{U}\mathbf{\Sigma}^2\mathbf{U}^\top
\end{aligned}
$$

- Observe that, $\mathbf{C}_\mathbf{x}$ is free of the terms $\mathbf{s}$ and $\mathbf{V}$.
- Since the covariance matrix $\mathbf{C}_\mathbf{x}$ is a symmetric matrix we can use the Spectral theorem to find $\mathbf{U}$ and $\mathbf{\Sigma}$.
- Let $\mathbf{C}_\mathbf{x} = \mathbf{E}\mathbf{D}\mathbf{E}^\top$ be the EVD of $\mathbf{C}_\mathbf{x}$.
- Then, we have $\mathbf{U} = \mathbf{E}$ and $\mathbf{\Sigma} = \mathbf{D}^{\frac{1}{2}}$.

Independent Component Analysis

- Therefore, we have that $\mathbf{W} = \mathbf{V}\boldsymbol{\Sigma}^{-1}\mathbf{U}^\top = \mathbf{V}\mathbf{D}^{-\frac{1}{2}}\mathbf{E}^\top$.
- Here, now $\mathbf{V}$ is the only unknown matrix.
- Now, the unmixed data becomes

$$
\begin{aligned}
\hat{\mathbf{s}} &= \mathbf{V}\mathbf{D}^{-\frac{1}{2}}\mathbf{E}^\top\mathbf{x} = \mathbf{V}\mathbf{x}_w, \ \text{here } \mathbf{x}_w = \mathbf{D}^{-\frac{1}{2}}\mathbf{E}^\top\mathbf{x} \\
\mathbb{E}\left[\hat{\mathbf{s}}\hat{\mathbf{s}}^\top\right] &= \mathbb{E}\left[\mathbf{V}\mathbf{D}^{-\frac{1}{2}}\mathbf{E}^\top\mathbf{x}\mathbf{x}^\top\mathbf{E}\mathbf{D}^{-\frac{1}{2}}\mathbf{V}^\top\right] \\
&= \mathbf{V}\mathbf{D}^{-\frac{1}{2}}\mathbf{E}^\top\mathbb{E}\left[\mathbf{x}\mathbf{x}^\top\right]\mathbf{E}\mathbf{D}^{-\frac{1}{2}}\mathbf{V}^\top \\
&= \mathbf{V}\mathbf{D}^{-\frac{1}{2}}\mathbf{E}^\top\mathbf{E}\mathbf{D}\mathbf{E}^\top\mathbf{E}\mathbf{D}^{-\frac{1}{2}}\mathbf{V}^\top \\
&= \mathbf{V}\mathbf{D}^{-\frac{1}{2}}\mathbf{D}\mathbf{D}^{-\frac{1}{2}}\mathbf{V}^\top \\
&= \mathbf{V}\mathbf{V}^\top \\
&= \mathbf{I} \\
\mathbb{E}\left[\mathbf{x}_w\mathbf{x}_w^\top\right] &= \mathbb{E}\left[\mathbf{D}^{-\frac{1}{2}}\mathbf{E}^\top\mathbf{x}\mathbf{x}^\top\mathbf{E}\mathbf{D}^{-\frac{1}{2}}\right] \\
&= \mathbf{D}^{-\frac{1}{2}}\mathbf{E}^\top\mathbb{E}\left[\mathbf{x}\mathbf{x}^\top\right]\mathbf{E}\mathbf{D}^{-\frac{1}{2}} \\
&= \mathbf{D}^{-\frac{1}{2}}\mathbf{E}^\top\mathbf{E}\mathbf{D}\mathbf{E}^\top\mathbf{E}\mathbf{D}^{-\frac{1}{2}} \\
&= \mathbf{D}^{-\frac{1}{2}}\mathbf{D}\mathbf{D}^{-\frac{1}{2}} \\
&= \mathbf{I}.
\end{aligned}
$$



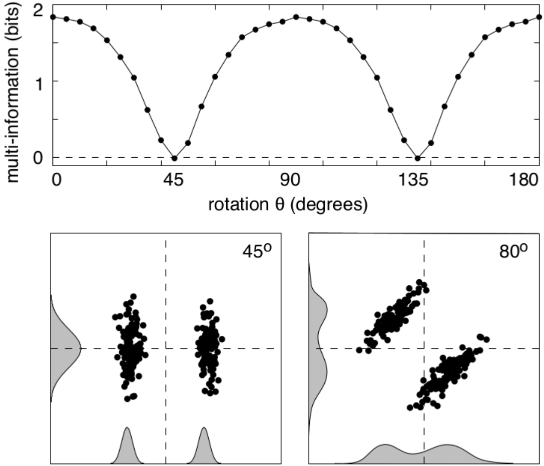
Independent Component Analysis

- Therefore, we have that $\mathbf{W} = \mathbf{V}\mathbf{D}^{-\frac{1}{2}}\mathbf{E}^\top$. Here, now $\mathbf{V}$ is the only unknown matrix.
- We, also have that: $\hat{\mathbf{s}} = \mathbf{V}\mathbf{x}_w$.
- We now exploit the statistical independence of the sound sources to find $\mathbf{V}$.
- We assume that all sources are statistically independent, thus: $P(\mathbf{s}) = \prod\limits_i p(s_i)$.
- Find optimal rotation $\mathbf{V}$ such that $\hat{\mathbf{s}}$ is statistically independent: $P(\hat{\mathbf{s}}) = \prod_i P(\hat{s}_i)$.
- **Mutual and Multi Information**: The mutual information measures the departure of two variables from statistical independence. The multi-information, a generalization of mutual information, measures the statistical dependence between multiple variables:

$$\begin{aligned}
I(\mathbf{s}) &= \int p(\mathbf{s}) \log_2 \left[\frac{p(\mathbf{s})}{\prod\limits_i p(s_i)} \right] d\mathbf{s} = \int p(\mathbf{s}) \log_2 (p(\mathbf{s})) d\mathbf{s} - \int p(\mathbf{s}) \log_2 \left(\prod_i p(s_i) \right) d\mathbf{s} \\
&= \int p(\mathbf{s}) \log_2 (p(\mathbf{s})) d\mathbf{s} - \sum_i \int p(s_i) \log_2 (p(s_i))\, ds_i \\
&= \sum_i \mathbb{H}[s_i] - \mathbb{H}[\mathbf{s}] = \sum_i \mathbb{H}\left[(\mathbf{V}\mathbf{x}_w)_i\right] - \mathbb{H}[\mathbf{V}\mathbf{x}_w] \\
&= \sum_i \mathbb{H}\left[(\mathbf{V}\mathbf{x}_w)_i\right] - \mathbb{H}[\mathbf{x}_w] - \log_2(|\mathbf{V}|) \\
\mathbf{V}^\star &= \operatorname*{arg\,min}_{\mathbf{V} \in \mathbb{R}^{2\times 2} : \mathbf{V}^\top \mathbf{V} = \mathbf{I}} \sum_i \mathbb{H}\left[(\mathbf{V}\mathbf{x}_w)_i\right]
\end{aligned}$$

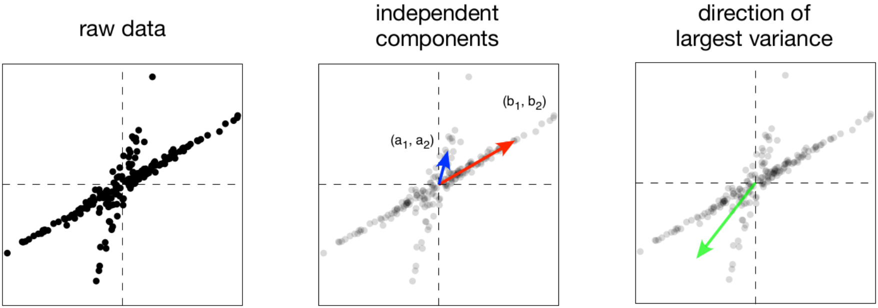
Independent Component Analysis

- The matrix $\mathbf{V}$ is a rotation matrix and in two dimensions it has the form $\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$
- The rotation angle θ is the only free variable.



ICA Algorithm

1: **Input**: $\mathbf{x}$
2: Subtract off the mean of the data in each dimension
3: $\mathbf{C_x} \leftarrow \mathbb{E}\left[\mathbf{x}\mathbf{x}^\top\right]$
4: $\mathbf{E}\mathbf{D}\mathbf{E}^\top \leftarrow \texttt{EVD}(\mathbf{C_x})$
5: $\mathbf{V}^\star \leftarrow \underset{\mathbf{V}\in\mathbb{R}^{2\times 2}:\mathbf{V}^\top\mathbf{V}=\mathbf{I}}{\arg\min} \sum_i \mathbb{H}\left[(\mathbf{V}\mathbf{x}_w)_i\right]$
6: $\mathbf{W} \leftarrow \mathbf{V}^\top\mathbf{D}^{-\frac{1}{2}}\mathbf{E}^\top$
7: $\hat{\mathbf{s}} \leftarrow \mathbf{W}\mathbf{x}$.



Linear Discriminant Analysis

- Consider the problem of predicting a label $y \in \{0, 1\}$ based on a feature vector $\mathbf{x} \in \mathbb{R}^d$.
- Now, using the Bayes rule we can write the optimal Bayes classifier as:

$$\begin{aligned} h_{\mathsf{Bayes}}(\mathbf{x}) &= \underset{y\in\{0,1\}}{\arg\max}\, p(y|\mathbf{x}) = \underset{y\in\{0,1\}}{\arg\max}\, \frac{p(y)p(\mathbf{x}|y)}{p(\mathbf{x})} \\ &= \underset{y\in\{0,1\}}{\arg\max}\, p(y)p(\mathbf{x}|y) \\ &= \arg\max\{p(y=1)p(\mathbf{x}|y=1), p(y=0)p(\mathbf{x}|=0)\} \end{aligned}$$

- Hence, $h_{\mathsf{Bayes}}(\mathbf{x}) = 1$ if and only if $p(y=1)p(\mathbf{x}|y=1) > p(y=0)p(\mathbf{x}|y=0)$.

$$\Rightarrow \frac{p(y=1)p(\mathbf{x}|y=1)}{p(y=0)p(\mathbf{x}|y=0)} > 1 \Rightarrow \log\left(\frac{p(y=1)p(\mathbf{x}|y=1)}{p(y=0)p(\mathbf{x}|y=0)}\right) > 0.$$

- We assume that $p(y=0) = p(y=1) = \frac{1}{2}$ and the conditional probability of X given Y is a Gaussian distribution.
- The covariance matrix of the Gaussian distribution is the same for both values of the label.
- Let $\boldsymbol{\mu}_0, \boldsymbol{\mu}_1 \in \mathbb{R}^d$ and let $\boldsymbol{\Sigma}$ be a covariance matrix. Then, the distribution is given by

$$p(\mathbf{x}|y) = \frac{1}{(2\pi)^{\frac{d}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_y)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_y)}.$$

Linear Discriminant Analysis

- This ratio is often called the log-likelihood ratio. In our case, the log-likelihood ratio becomes

$$
\begin{aligned}
\log\left(\frac{p(y=1)p(\mathbf{x}|y=1)}{p(y=0)p(\mathbf{x}|y=0)}\right) &= \log\left(\frac{e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)}}{e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_0)^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_0)}}\right) \\
&= \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_0)^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_0) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) \\
&= \mathbf{x}^\top(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)^\top\boldsymbol{\Sigma}^{-1} + \frac{1}{2}\left(\boldsymbol{\mu}_0^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1\right) \\
&= \mathbf{x}^\top\boldsymbol{\omega} + b.
\end{aligned}
$$

- Therefore, $y = 1$ if $\mathbf{x}^\top\boldsymbol{\omega} + b > 0$ and $y = 0$ if $\mathbf{x}^\top\boldsymbol{\omega} + b < 0$. Here, $\mathbf{x}^\top\boldsymbol{\omega} + b = 0$ is the hyperplane separating two classes.
- Under these generative assumptions, the Bayes optimal classifier is a linear classifier.
- Additionally, one may train the classifier by estimating the parameter $\boldsymbol{\mu}_0$, $\boldsymbol{\mu}_1$, $\boldsymbol{\Sigma}$ from the data, using the maximum likelihood estimator.
- With those estimators at hand, the values of $\mathbf{w}$ and b can be calculated as above.