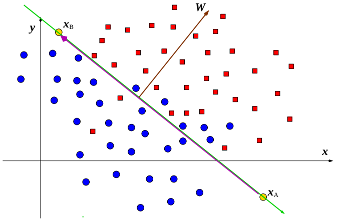
Classification

Juan Neirotti

October 12, 2022

Discriminants

- Discriminants are functions used to separate elements of a set into classes.
- Therefore, the data set is formed by order pairs of the form $(t_n, \boldsymbol{x}_n)$, where $t_n \in \{-1, +1\}$ and $\boldsymbol{x}_n \in \mathbb{R}^d$.
- Discriminants depend on parameters that can be adjusted by optimizing a given cost function.
- The simplest discriminant function consist of a linear combination of the input variables, in which the coefficients of the linear combination are the parameters of the model.
- Consider for instance the straight line $y(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{x} + w_0$ where $\boldsymbol{w}$ and w_0 have been chosen to maximize the perpendicular distance between the line $y(\boldsymbol{x})$ and the data-points shown in the following plot.



Inference and Discriminant Functions

- We start by considering a two-category classification problem. The joint probability of the labels and features variables is $\mathcal{P}(t, \boldsymbol{x})$. Then:

$$
\begin{aligned}
\mathcal{P}(\boldsymbol{x}) &= \sum_{t=-1,+1} \mathcal{P}(t, \boldsymbol{x}) \\
\mathcal{P}(t) &= \int \mathrm{d}\boldsymbol{x} \mathcal{P}(t, \boldsymbol{x}) \\
\mathcal{P}(t|\boldsymbol{x}) &= \frac{\mathcal{P}(t, \boldsymbol{x})}{\mathcal{P}(\boldsymbol{x})} \\
&= \frac{\mathcal{P}(\boldsymbol{x}|t)\mathcal{P}(t)}{\sum_{t=-1,+1} \mathcal{P}(\boldsymbol{x}|t)\mathcal{P}(t)}.
\end{aligned}
$$

- We usually model $\mathcal{P}(\boldsymbol{x}|t)$ (the likelihood of the class t) in order to compute the posterior probability of class t ($\mathcal{P}(t|\boldsymbol{x})$).
- The *boundary* between the classes is the region $\boldsymbol{x}^\star \in \mathbb{R}^d$ such that $\mathcal{P}(t = -1|\boldsymbol{x}^\star) = \mathcal{P}(t = +1|\boldsymbol{x}^\star)$.

Discriminant Functions

- Observe that by defining the functions

$$y_{+1}(\boldsymbol{x}) = \ln\left(\frac{\mathcal{P}(\boldsymbol{x}|+1)\mathcal{P}(+1)}{\mathcal{P}(\boldsymbol{x})}\right)$$

$$y_{-1}(\boldsymbol{x}) = \ln\left(\frac{\mathcal{P}(\boldsymbol{x}|-1)\mathcal{P}(-1)}{\mathcal{P}(\boldsymbol{x})}\right)$$

$$\begin{aligned} y(\boldsymbol{x}) &= y_{+1}(\boldsymbol{x}) - y_{-1}(\boldsymbol{x}) \\ &= \ln\left(\mathcal{P}(\boldsymbol{x}|+1)\right) - \ln\left(\mathcal{P}(\boldsymbol{x}|-1)\right) + \ln\left(\frac{\mathcal{P}(+1)}{1-\mathcal{P}(+1)}\right) \end{aligned}$$

the boundary satisfies the equation $y(\boldsymbol{x}^\star) = 0$, the elements $\boldsymbol{x} \in \mathbb{R}^d$ that belong to the class with label +1 (-1) satisfy $y(\boldsymbol{x}) > 0 (< 0)$.

Gaussian Discriminant Functions

- If the priors $\mathcal{P}(t=+1)$ and $\mathcal{P}(t=-1)$ are given (by counting how many elements belong to each class for instance) and the likelihoods are modeled by Gaussian distributions:

$$\mathcal{P}(\boldsymbol{x}|t) = \frac{1}{\sqrt{2\pi|\boldsymbol{\Sigma}_t|}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_t)^{\mathrm{T}}\boldsymbol{\Sigma}_t^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_t)\right\} \quad (1)$$

  where $\boldsymbol{\mu}_t \in \mathbb{R}^d$ is the center and $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ is the covariance matrix of the class-t Gaussian.

- The discriminant function becomes:

$$y(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_{+1})^{\mathrm{T}}\boldsymbol{\Sigma}_{+1}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_{+1}) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_{-1})^{\mathrm{T}}\boldsymbol{\Sigma}_{-1}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_{-1}) + \frac{1}{2}\ln\frac{|\boldsymbol{\Sigma}_{-1}|}{|\boldsymbol{\Sigma}_{+1}|} + \ln\left(\frac{\mathcal{P}(+1)}{1-\mathcal{P}(+1)}\right)$$

Gaussian Discriminant Functions

- It may occur that both classes have been generated by similar means which implies that $\boldsymbol{\Sigma}_{-1} = \boldsymbol{\Sigma}_{+1} = \boldsymbol{\Sigma}$, thus:

$$y(\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_{+1}-\boldsymbol{\mu}_{-1})-\frac{1}{2}\left(\boldsymbol{\mu}_{+1}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{+1} - \boldsymbol{\mu}_{-1}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{-1}\right),$$

  which is a polynomial of order 1 in $\boldsymbol{x}$ of the form $y(\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{w} + w_0$.
- In such a case we have that

$$\begin{aligned} \boldsymbol{w} &= \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_{+1} - \boldsymbol{\mu}_{-1}) \\ w_0 &= -\frac{1}{2}\left(\boldsymbol{\mu}_{+1}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{+1} - \boldsymbol{\mu}_{-1}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{-1}\right). \end{aligned}$$

- The boundary is defined by the equation of the plane:

$$\boldsymbol{x}^{\mathrm{T}}\boldsymbol{w} = -w_0.$$

Logistic discriminant

- Let us consider a network with non-linear output:

$$y = g(a)$$
$$a = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{w} + w_0.$$

- For the two-class problem with likelihood given by (1) with equal covariance for both classes, we have that:

$$\mathcal{P}(t=+1|\boldsymbol{x}) = \frac{\mathcal{P}(\boldsymbol{x}|t=+1)\mathcal{P}(t=+1)}{\mathcal{P}(\boldsymbol{x}|t=+1)\mathcal{P}(t=+1)+\mathcal{P}(\boldsymbol{x}|t=-1)\mathcal{P}(t=-1)}$$
$$= \frac{1}{1+\exp(-a)} = g(a)$$

where

$$a = \ln \frac{\mathcal{P}(\boldsymbol{x}|t=+1)\mathcal{P}(t=+1)}{\mathcal{P}(\boldsymbol{x}|t=-1)\mathcal{P}(t=-1)}$$

and

$$\boldsymbol{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_{+1} - \boldsymbol{\mu}_{-1})$$
$$w_0 = -\frac{1}{2}\left(\boldsymbol{\mu}_{+1}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{+1} - \boldsymbol{\mu}_{-1}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_{-1}\right) + \ln\frac{\mathcal{P}(t=+1)}{\mathcal{P}(t=-1)}.$$

K-Means Algorithm

- Suppose our data set is composed by no-labeled data $\mathcal{D} = \{\boldsymbol{x}_n\}$.
- The K-means algorithm assumes one is given the data set $\mathcal{D}$ with the goal of partitioning the N observations into $k \ll N$ classes.
- The K is unknown.
- The k-means algorithm can be described as the following optimization algorithm:

$$\underset{\boldsymbol{\mu}_j}{\operatorname{argmin}} \sum_{j=1}^{K} \sum_{\boldsymbol{x}_n \in \mathcal{D}_j} \left\| \boldsymbol{x}_n - \boldsymbol{\mu}_j \right\|^2,$$

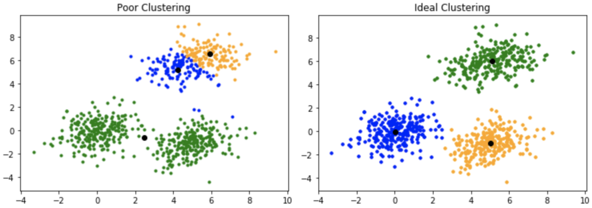
  where $\boldsymbol{\mu}_j$ is the mean of the jth cluster and $\mathcal{D}_j$ is the subset of data points in this cluster.
- Solving the optimization problem is NP-hard (we will come back to the discussion of algorithm complexity as soon as possible).

K-Means Algorithm- Pseudo Algorithm

1. Select (*by an appropriate way*) the K initial prototypes $\{\boldsymbol{\mu}_k\}_{k=1}^{K}$.
2. Repeat until convergence
    2.1 Measure the distance of all the points to the K prototypes $\{\boldsymbol{\mu}_k\}$.
    2.2 Assign to class k the points that are closer to $\boldsymbol{\mu}_k$ than to the other prototypes $\boldsymbol{\mu}_{j\neq k}$.
    2.3 Compute the new prototypes by averaging over the clusters formed in the previous step.

K-Means Algorithm- Problems

- Depending on the way the initialization is done, we may encounter problems

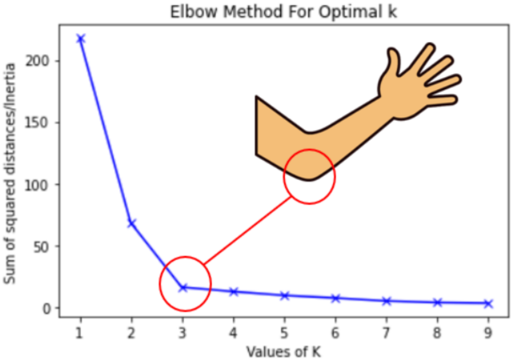


K-Means++ Algorithm

1. Select the first prototype at random from the data set $\boldsymbol{\mu}_1$.
2. Find the point of the data set that is farthest away from the first prototype and choose it as $\boldsymbol{\mu}_2$.
3. Repeat until finding the remaining $k-2$ prototypes
   3.1 For each point $\boldsymbol{x}_n \in \mathcal{D}$ find $d_{n,k} = \|\boldsymbol{x}_n - \boldsymbol{\mu}_k\|$ and store $d_n = \min\{d_{n,k}\}$.
   3.2 For $m = \mathrm{argmax}_n\{d_n\}$, assign the next prototype to $\boldsymbol{x}_m = \boldsymbol{\mu}_{k+1}$.

K-Means Algorithm- Model Selection

- Elbow Method: Plot the mean square error vs k. Not very reliable (we will use it during the lab session).



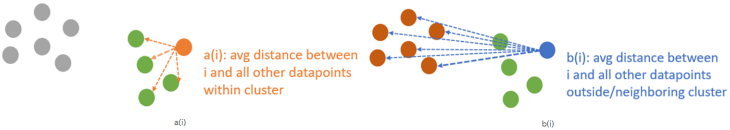
Line plot between K and inertia

K-Means Algorithm- Model Selection

- Silhouette Method: Makes use of measures of similarity and dissimilarity::

$$a_i = \frac{1}{|C_i| - 1} \sum_{j \in C_i, j \neq i} \|\mathbf{x}_i - \mathbf{x}_j\|$$

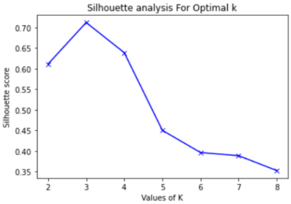
$$b_i = \min_{i \neq j} \frac{1}{|C_j|} \sum_{j \in C_j} \|\mathbf{x}_i - \mathbf{x}_j\|$$



K-Means Algorithm- Model Selection

- Silhouette Method: Makes use of measures of similarity and dissimilarity::

$$s_i = \frac{b_i - a_i}{\max\{a_i, b_i\}}$$



Line plot between K and Silhouette score