Linear Discriminant Analysis

Jack Yang

October 14, 2020

Outline

Classification

What is Classification?

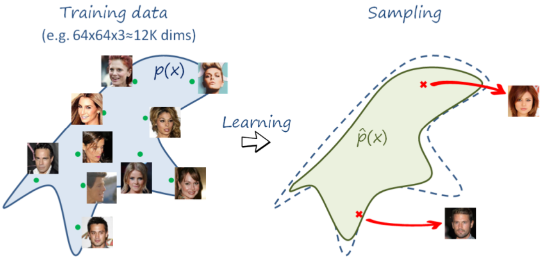
Classification is assigning a d-dimensional data point to one of a discrete number of classes.

Building a Classifier

- Generative Models (e.g., Linear Discriminant Analysis)
- Discriminative Models (e.g., Logistic Regression)

Generative Models

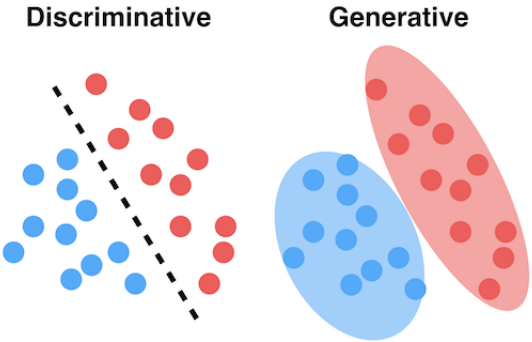
Given the training data, we want to generate new samples from same distribution.



Assume Training Data $\sim p_{\text{data}}(x)$ and Generating Data $\sim p_{\text{model}}(x)$, we want to learn Generating Data $\sim p_{\text{model}}(x)$ similar to Training Data $\sim p_{\text{data}}(x)$.

Discriminative Models

Discriminative models directly learn a decision boundary.



Linear Discriminant Analysis

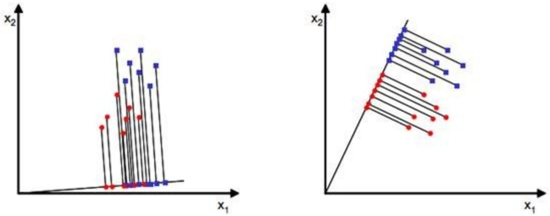
Intuitively, a good classifier is one that bunches together observations in the same class and separates observations between classes.

- Fisher's Linear Discriminant attempts to do this through dimensionality reduction.
  - Specifically, it projects data points onto a single dimension and classifies them according to their location along this dimension.
- Bayes' Linear Discriminant attempts to do this through probabilistic modeling.

Fisher's Linear Discriminant

Separate samples of distinct groups by projecting them onto a space that

- Maximizes their between-class separability while
- Minimizing their within-class variability



Approach: Maximize Class Separation

Assume we have d-dimensional samples $\{x_1, x_2, \ldots, x_N\}$. We seek to obtain a scalar y by projecting the samples x onto a line.

Consider two classes: C_1 with N_1 points and C_2 with N_2 points.

- The corresponding mean vectors

$$\mu_1 = \frac{1}{N_1} \sum_{n \in C_1} x_n \qquad \mu_2 = \frac{1}{N_2} \sum_{n \in C_2} x_n.$$

- Measure class separation as the distance of the projected class

$$|\tilde{\mu_2} - \tilde{\mu_1}| = \left| w^T \mu_2 - w^T \mu_1 \right| = \left| w^T (\mu_2 - \mu_1) \right|$$

where w is the projection vectors used to project x to y and we want to maximize this with respect to w with the constraint $\|w\| = 1$.

Approach: Minimize Within-Class Variability

We want to maximize their between-class separability **and** minimize their within-class variablity.

For each class C_k, the **within-class scatter** is given as

$$s_k = \sum_{n \in C_k} (y_n - \tilde{\mu_k})^2$$

where $y_n = w^T x_n$ and $\tilde{\mu_k} = w^T \mu_k$.

Maximize Fisher Criterion

Let

$$J(w) = \frac{\text{Between-Class Scatter}}{\text{Within-Class Scatter}} = \frac{(\tilde{\mu_2} - \tilde{\mu_1})^2}{s_1^2 + s_2^2} = \frac{w^T S_B w}{w^T S_W w}$$

where

$$S_B = (\mu_2 - \mu_1)(\mu_2 - \mu_1)^T, \quad S_W = \sum_k \sum_{n \in C_k} (x_n - \mu_k)(x_n - \mu_k)^T.$$

By Lagrange Multiplier, we will finally get

$$w = S_W^{-1}(\hat{\mu_2} - \hat{\mu_1}).$$

Naive Bayes

Naive Bayes, is a classifier based on Bayes Theorem with the "naive" assumption that features are independent of each other.

Theorem (Bayes' Theorem)

Given a feature vector $X = (x_1, x_2, \ldots, x_n)$ *and a class variable* C_k*,*

$$\mathbb{P}(C_k|X) = \frac{\mathbb{P}(X|C_k)\mathbb{P}(C_k)}{\mathbb{P}(X)}$$

for $k = 1, 2, \ldots, K$*.*
We call $\mathbb{P}(C_k|X)$ *the posterior probability,* $\mathbb{P}(X|C_k)$ *the likelihood,* $\mathbb{P}(C_k)$ *the prior probability of class, and* $\mathbb{P}(X)$ *the prior probability of predictor.*

Naive Independence Assumption

The likelihood function can be decomposed as

$$\begin{aligned}\mathbb{P}(X|C_k) &= \mathbb{P}(x_1, x_2, \ldots, x_n|C_k)\\ &= \mathbb{P}(x_1|x_2, \ldots, x_n, C_k)\cdot\mathbb{P}(x_2|x_3, \ldots, x_n, C_k)\cdots\mathbb{P}(x_{n-1}, x_n|C_k)\cdot\mathbb{P}(x_n|C_k).\end{aligned}$$

With the naive conditional independence assumption, that is,

$$\mathbb{P}(x_i|x_{i+1}, \ldots, x_n, C_k) = \mathbb{P}(x_i|C_k).$$

Now, the likelihood is $\mathbb{P}(X|C_k) = \displaystyle\prod_{i=1}^{n}\mathbb{P}(x_i|C_k)$. Therefore, the posterior probability can be evaluated as

$$\mathbb{P}(C_k|X) = \frac{\mathbb{P}(C_k)}{\mathbb{P}(X)}\cdot\prod_{i=1}^{n}\mathbb{P}(x_i|C_k).$$

Naive Bayes Model

Since the prior probability of predictor $\mathbb{P}(X)$ is constant, we can get the following proportional relation

$$\mathbb{P}(C_k|X) \propto \mathbb{P}(C_k) \prod_{i=1}^{n} \mathbb{P}(x_i|C_k).$$

Now, we want to find a class $\hat{C}$ that maximizes the posterior probability,

$$\hat{C} = \arg\max_{C_k} \mathbb{P}(C_k) \prod_{i=1}^{n} \mathbb{P}(x_i|C_k).$$

Further, we may use Maximum Likelihood Estimate to find which class it should be.