Bayes Classifier
and
Linear Discriminant Analysis

Announcements

- All waitlisted students have been granted permission to enroll.
- HW 1 will go out today, and is due in one week.
- There will be a Python tutorial next week, stayed tuned for details.
- Please use Piazza for course related questions.
- Office hours have been posted to Canvas

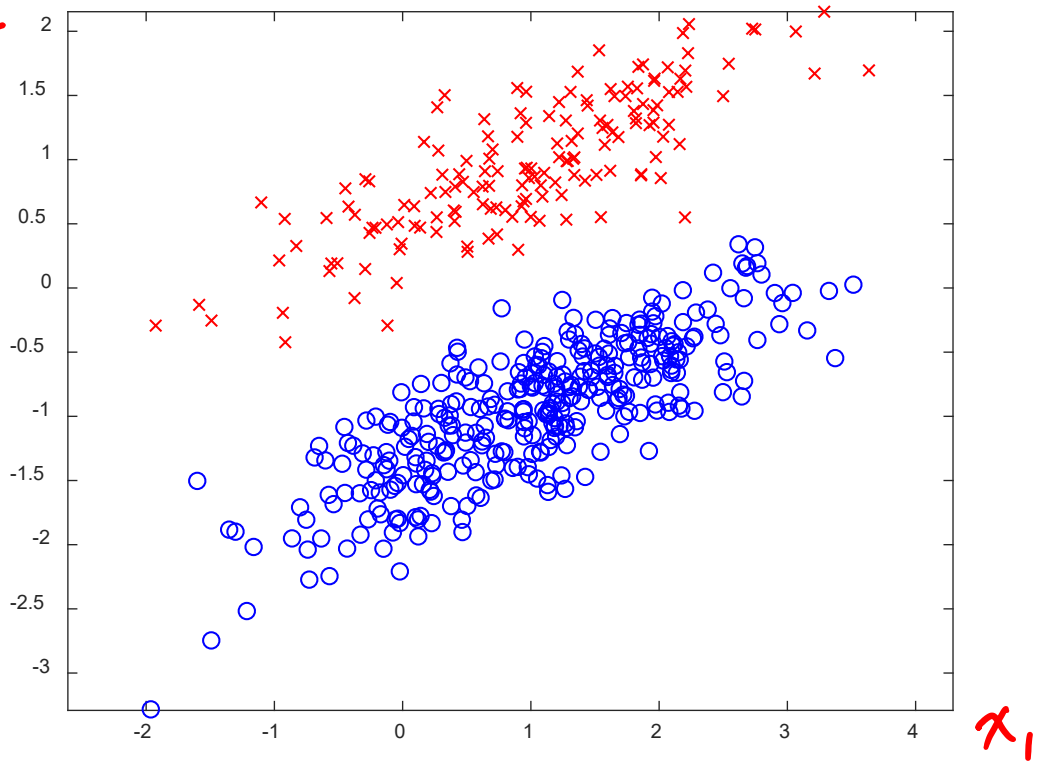
Classification: Probabilistic Setting

- Feature vector $\boldsymbol{X} \in \mathbb{R}^d$
- Label $Y \in \{1, \ldots, K\}$
- Assume $(\boldsymbol{X}, Y)$ are jointly distributed ($d+1$ dimensional)

Y is the label associated to X

$(x_1, y_1), \ldots, (x_n, y_n)$ iid joint distrib. of (x, y)

Classification: Probabilistic Setting



$\pi_1 = \frac{1}{3}$

$\pi_2 = \frac{2}{3}$

$x \in \mathbb{R}^2$

$y \in \{1, 2\}$

$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

(x, y)

- Two classes, each class has a Gaussian distribution
- Blue class is twice as probable as red class
- How can we characterize the joint distribution mathematically?

Classification: Probabilistic Setting

- How should we think about the joint distribution of $(\boldsymbol{X}, Y)$?
- First way: Marginal distribution of Y, and conditional distribution of $\boldsymbol{X}$ given $Y = y$, for each y
- Second way: Marginal distribution of $\boldsymbol{X}$, and conditional distribution of Y given $\boldsymbol{X} = \boldsymbol{x}$, for each $\boldsymbol{x}$

Classification: Probabilistic Setting

- Notation:
  - $\pi_k := \Pr(Y = k)$ — class prior
  - $g_k(\boldsymbol{x}) :=$ pdf/pmf of $\boldsymbol{X}$ given $Y = k$ — class-conditional pdf/pmf

  (1st way)
  - $\eta_k(\boldsymbol{x}) := \Pr(Y = k | \boldsymbol{X} = \boldsymbol{x})$ — class posterior
  - $g(\boldsymbol{x}) :=$ pdf/pmf of $\boldsymbol{X}$ — marginal distrib. of X

  (2nd way)

$$g(x) = \sum_{k=1}^{K} \pi_k g_k(x)$$

$$\eta_k(x) = \frac{\pi_k g_k(x)}{g(x)}$$

Classification: Probabilistic Setting

- First way: Marginal distribution of Y, and conditional distribution of $\boldsymbol{X}$ given $Y = y$, for each y

red class : $y = 1$

blue " : $y = 2$

$\pi_1 = \frac{1}{3}$, $\pi_2 = \frac{2}{3}$

$X \mid Y=1 \sim N(\mu_1, \Sigma)$

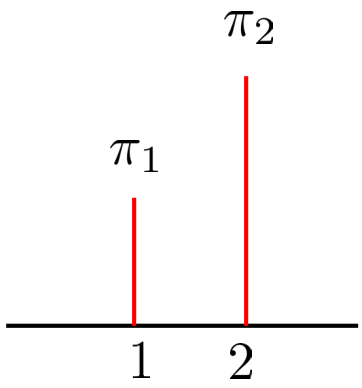
$X \mid Y=2 \sim N(\mu_2, \Sigma)$

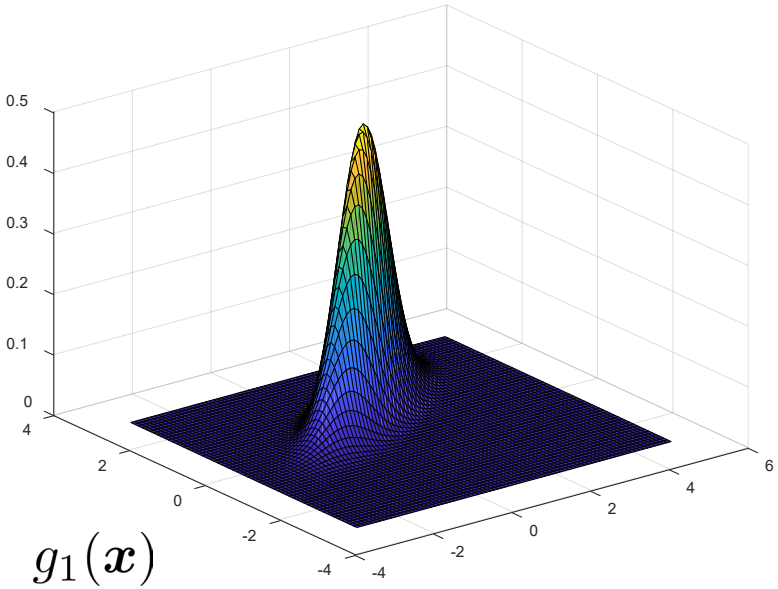
$$\phi(x; \mu, \Sigma) = (2\pi)^{-\frac{d}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right\}$$

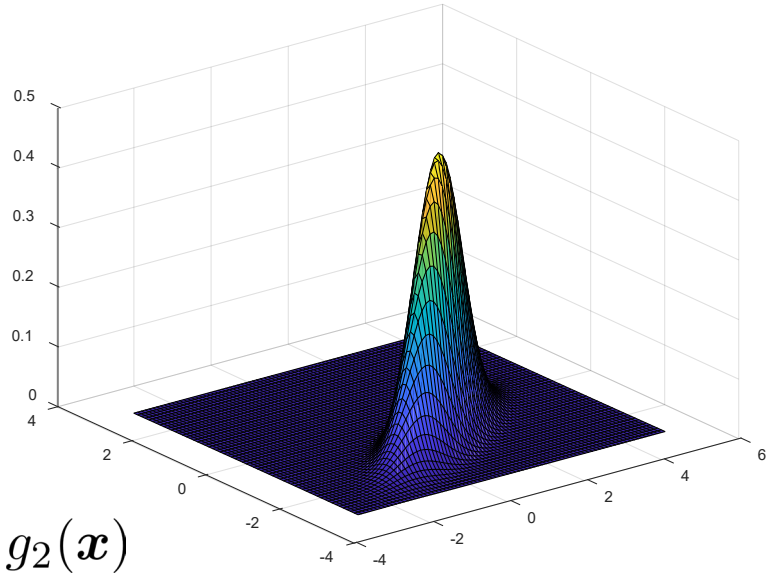
$$g_1(x) = \phi(x; \mu_1, \Sigma)$$

$$g_2(x) = \phi(x; \mu_2, \Sigma)$$

Classification: Probabilistic Setting







Classification: Probabilistic Setting

- Second way: Marginal distribution of $\boldsymbol{X}$, and conditional distribution of Y given $\boldsymbol{X} = \boldsymbol{x}$, for each $\boldsymbol{x}$

$$g(x) = \pi_1 g_1(x) + \pi_2 g_2(x)$$

$$= \frac{1}{3}\phi(x; \mu_1, \Sigma) + \frac{2}{3}\phi(x; \mu_2, \Sigma)$$

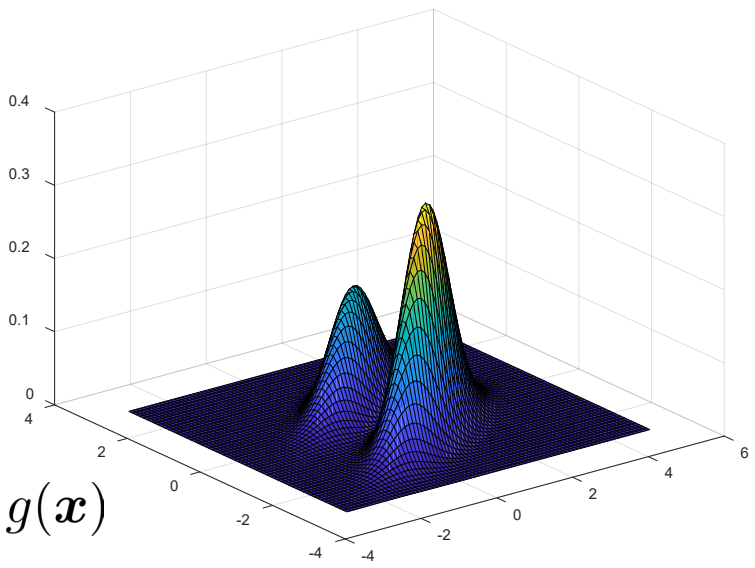
$$\eta_1(x) = \frac{\pi_1 g_1(x)}{\pi_1 g_1(x) + \pi_2 g_2(x)}$$

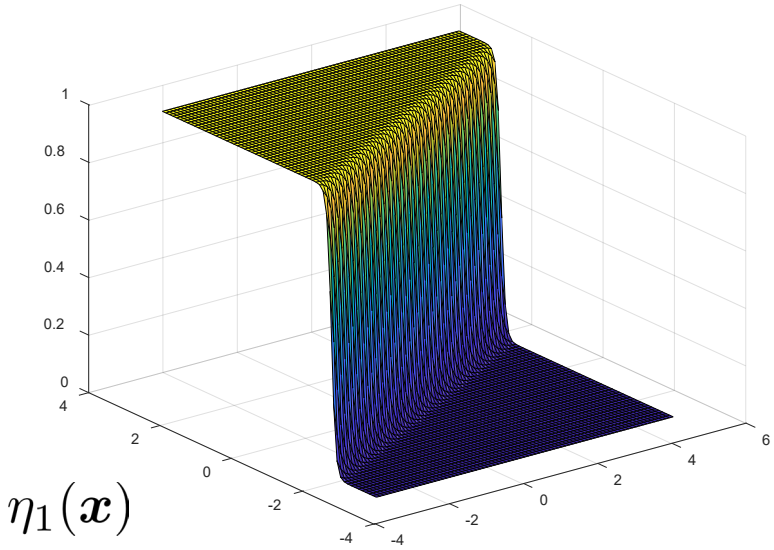
$$\eta_2(x) = 1 - \eta_1(x)$$

Classification: Probabilistic Setting

$$g(\boldsymbol{x}) = \pi_1 g_1(\boldsymbol{x}) + \pi_2 g_2(\boldsymbol{x})$$

$$\eta_1(\boldsymbol{x}) = \frac{\pi_1 g_1(\boldsymbol{x})}{\pi_1 g_1(\boldsymbol{x}) + \pi_2 g_2(\boldsymbol{x})}$$





Bayes Classifier

- A *classifier* is a function $f : \mathbb{R}^d \to \{1, \ldots, K\}$
- Given a joint distribution of $(\boldsymbol{X}, Y)$, what is the best possible classifier?
- The best classifier depends on the performance measure. The most common performance measure is the probability of error, or *risk*, defined by:

$$R(f) = \Pr(f(X) \neq Y) \quad \in [0,1]$$

  i.e., the probability of the event

$$\{(\boldsymbol{x}, y) \in \mathbb{R}^d \times \{1, \ldots, K\} \mid f(\boldsymbol{x}) \neq y\}$$

  sample space

- The smallest risk of any classifier is called the Bayes risk and is denoted R^*
- If $R(f) = R^*$, f is called a Bayes classifier

Bayes Classifier

- **Theorem:** The classifier

$$f^*(\boldsymbol{x}) = \underset{k \in \{1,\dots,K\}}{\arg\max}\ \eta_k(x)$$

$$= \underset{k \in \{1,\dots,K\}}{\arg\max}\ \pi_k g_k(x)$$

is a Bayes classifier.

$\eta_k(x) = \Pr(Y = k \mid X = x)$

The pmf of $Y \mid X = x$ is $\eta_k(x)$

$$\eta_k(x) = \frac{\pi_k g_k(x)}{g(x)}$$

Poll

True or false: A Bayes classifier always predicts the correct label

(A) True

(B) False

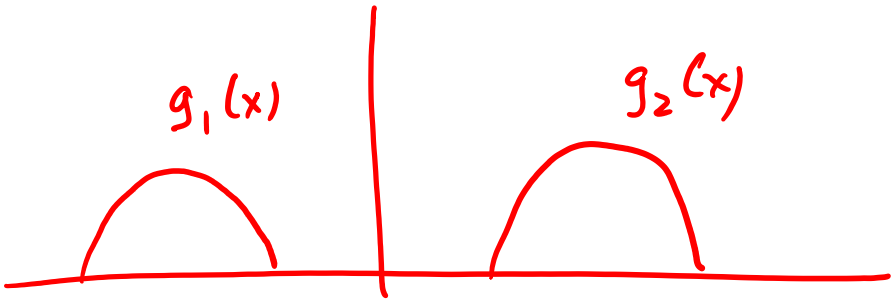
Questions

1. Give an example of a classification problem where $R^* = 0$.
2. Give an example of a classification problem where $R^* = \frac{1}{2}$.
3. What is R^* when $f(\boldsymbol{x})$ always predicts class k?

1. $\pi_1 = 1, \pi_2 = 0$

or

$g_1(x)$ and $g_2(x)$ don't overlap



2. $g_1(x) = g_2(x)\ \forall x$

$\eta_1(x) = \frac{1}{2}\ \forall x$

3. $1 - \pi_k$

Plug-In Classifiers

- In machine learning, the quantities $\pi_k, g_k(\boldsymbol{x})$, $\eta_k(\boldsymbol{x})$, are not known, so we can't know the Bayes' classifier.
- However, the formula for the Bayes' classifier is still useful. We can estimate the quantities in the formula from training data

$$(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n),$$

  and plug those estimates in to the formula to get a classifier.
- Linear discriminant analysis and Naïve Bayes have the form

$$f(\boldsymbol{x}) := \arg\max_k \widehat{\pi}_k \widehat{g}_k(\boldsymbol{x})$$

- Logistic regression has the form

$$f(\boldsymbol{x}) := \arg\max_k \widehat{\eta}_k(\boldsymbol{x})$$

- Other estimators are possible.

Linear Discriminant Analysis

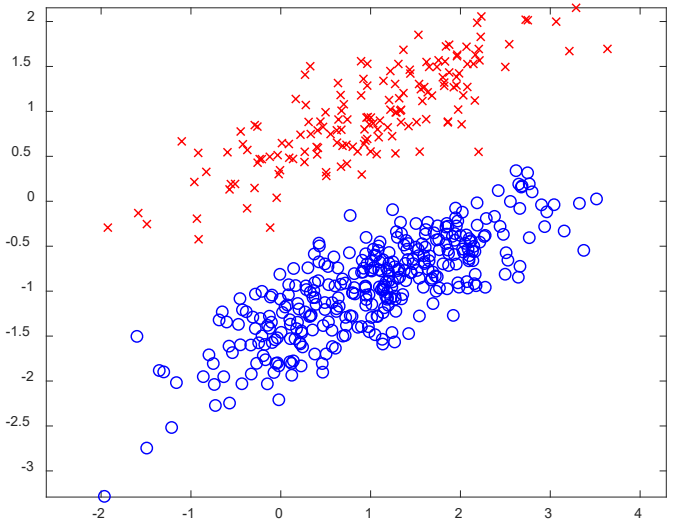
- Training data $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$.
- *LDA assumption*:

$$y_1, \ldots, y_n \overset{iid}{\sim} \text{discrete}(\pi_1, \ldots, \pi_K)$$
$$\boldsymbol{x} \mid y = k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}), \quad k = 1, \ldots, K$$

for some unknown $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$ and $\boldsymbol{\Sigma}$. Equivalently

$$g_k(\boldsymbol{x}) = (2\pi)^{-d/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k) \right)$$

- LDA is the plug-in rule based on this model. We use training data to estimate π_k, $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}$.



LDA Estimates

- LDA is the classifier obtained by plugging the following into the Bayes classifier formula:

$$\widehat{\pi}_k = \frac{n_k}{n}, \qquad n_k = |\{i : y_i = k\}|$$

$$\widehat{\boldsymbol{\mu}}_k = \frac{1}{n_k} \sum_{i:y_i=k} \boldsymbol{x}_i$$

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_{y_i})(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_{y_i})^T$$

$$\hat{f}(x) = \underset{k}{\operatorname{argmax}}\ \hat{\pi}_k \cdot \hat{g}_k(x)$$

$$= \underset{k}{\operatorname{argmax}}\ \hat{\pi}_k \cdot \phi(x; \hat{\mu}_k, \hat{\Sigma})$$

- These estimates are all *maximum likelihood estimates*
- $\widehat{\boldsymbol{\mu}}_k$ is the *sample mean* for each class
- Alternatively, $\widehat{\boldsymbol{\Sigma}}$ can be the *pooled sample covariance*,

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n-d} \sum_{i=1}^{n} (\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_{y_i})(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_{y_i})^T$$

LDA is a Linear Classifier

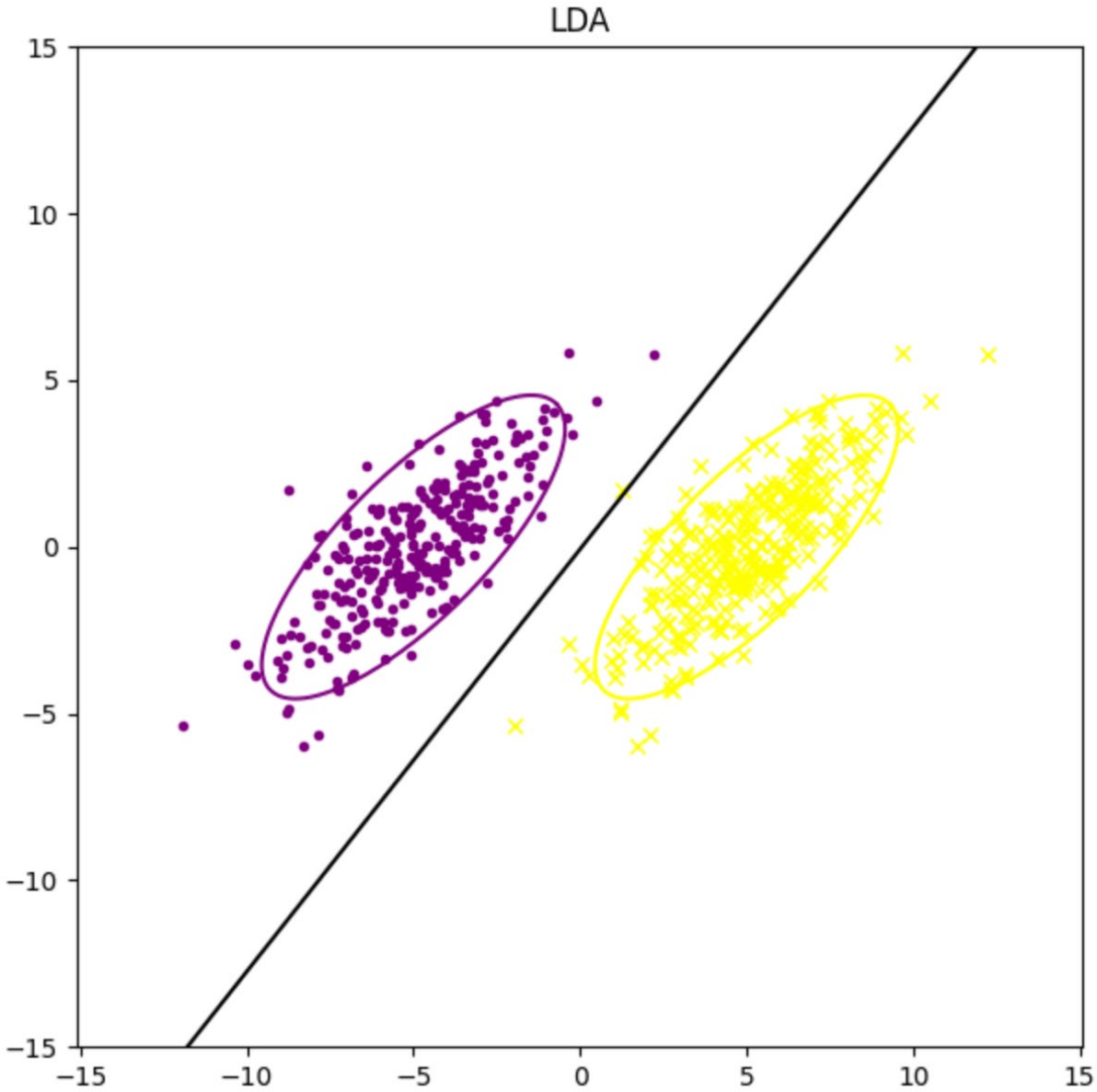
- Binary setting, $Y \in \{-1, 1\}$.
- The LDA classifier is

$$\widehat{f}(\boldsymbol{x}) = \text{sign}(\boldsymbol{w}^T \boldsymbol{x} + b)$$

where

$$\boldsymbol{w} = \widehat{\boldsymbol{\Sigma}}^{-1}(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_{-1})$$

$$b = \log(\widehat{\pi}_1/\widehat{\pi}_{-1}) + \frac{1}{2}(\widehat{\boldsymbol{\mu}}_{-1}^T \widehat{\boldsymbol{\Sigma}}^{-1} \widehat{\boldsymbol{\mu}}_{-1} - \widehat{\boldsymbol{\mu}}_1^T \widehat{\boldsymbol{\Sigma}}^{-1} \widehat{\boldsymbol{\mu}}_1).$$


LDA
15
10
5
0
−5
−10
−15
−15
−10
−5
0
5
10
15

Generative vs Discriminative

- A machine learning algorithm is *generative* if it is based on a full probabilistic model for the data, in other words, a model that could be used to generate data (without needing to have seen data previously).
- A machine learning algorithm is *discriminative* if it is not generative.
- LDA is generative

Parametric vs. Nonparametric

A learning algorithm is *nonparametric* if the amount of space needed to store the learned model (the case of LDA, a classifier) grows with n, otherwise it is *parametric*.

Poll: Is LDA parametric or nonparametric?

A. parametric ✓

B. nonparametric

LDA Summary

- One of the very earliest ML algorithms invented (c. 1930s)
- Many applications throughout history, although better methods now available
- Requires n > d although extensions are possible
- Logistic regression almost always preferable (makes a weaker assumption)
- Optional: read about Mahalanobis distance in lecture notes