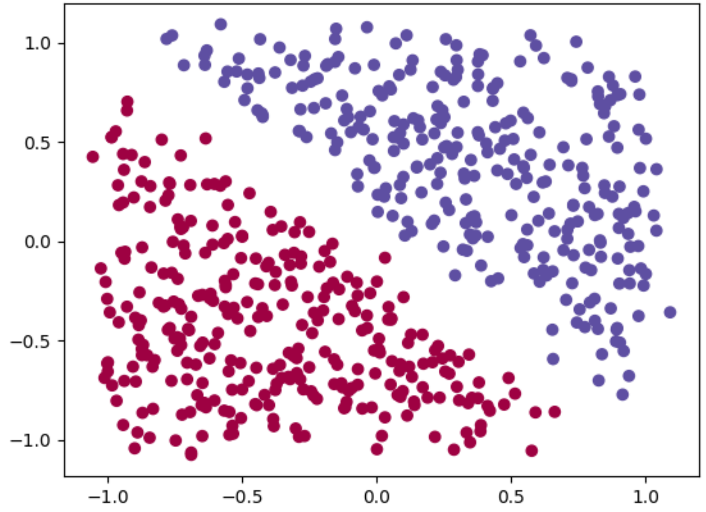
CSE 5523:
HW2

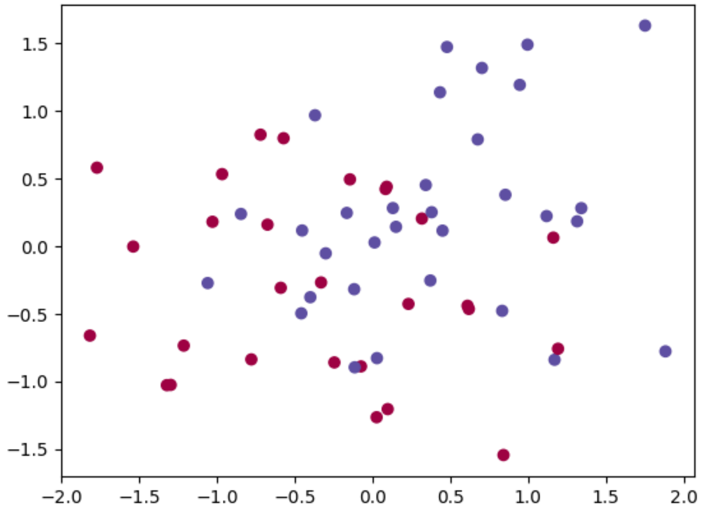
Outline

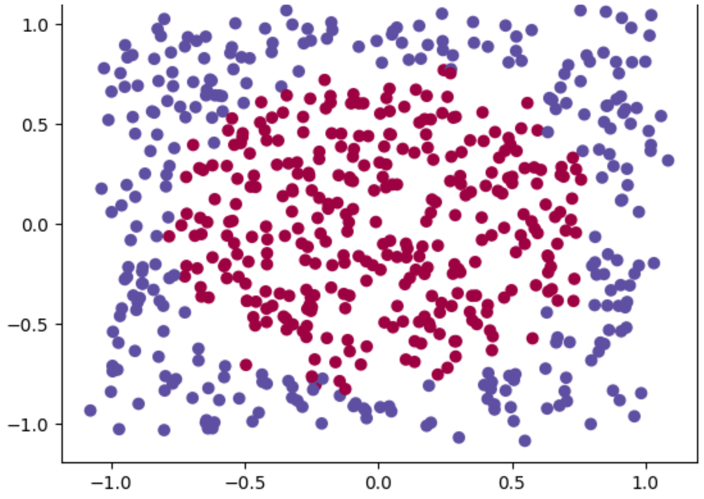
- You are to implement:
  - Pocket algorithm (improved perceptron)
  - Linear Gaussian discriminative analysis
  - Nonlinear Gaussian discriminative analysis

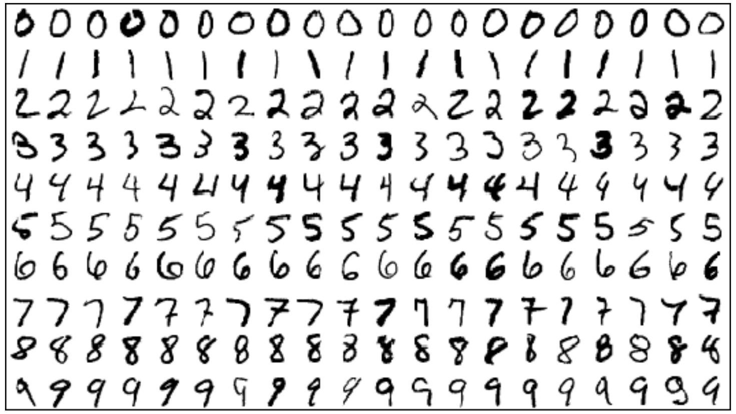
Data

- Four data source
  - ○ 2D linear
  - ○ 2D noisy linear
  - ○ 2D quadratic (circle)
  - ○ MNIST (<5 vs. >=5)
- $\boldsymbol{X} \in \mathbb{R}^{D \times N}$:
  - ○ A column as an instance
- $\boldsymbol{Y} \in \{+1, -1\}^{N \times 1}$









Data

- The data $\boldsymbol{X}$ are not appended with "1" yet.
- For feature transform for a 2D data instance $\boldsymbol{x} \in \mathbb{R}^2$, we do
  - $\boldsymbol{\phi}(\boldsymbol{x}) = \begin{bmatrix} x[1] \\ x[2] \\ x[1]^2 \\ x[2]^2 \\ x[1] \times x[2] \end{bmatrix}$
  - Again, you need to append "1" to the data $\boldsymbol{\phi}(\boldsymbol{x})$ if you want to solve $\tilde{\boldsymbol{w}}$ directly
  - In the homework, we have done $\boldsymbol{\phi}(\boldsymbol{x})$ for you!

Accuracy

- Data = $\{\ (\boldsymbol{x}_i,\ \ y_i \in \{+1, -1\})\ \}_{i=1}^{N}$
- Accuracy = $\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[\hat{y}_i\ ==\ y_i]$, where $\hat{y}_i$ is the prediction based on $\boldsymbol{x}_i$

Pocket algorithm

Pocket algorithm

- **Training data:** $D_{tr} = \{\left(\boldsymbol{x}_i \in \mathbb{R}^D,\ \mathrm{y}_i \in \{+1, -1\}\right)\}_{i=1}^{N}$

- **Model:** $\mathrm{sign}(\boldsymbol{w}\boldsymbol{x} + b) = \mathrm{sign}\left(\widetilde{\boldsymbol{w}}^T \widetilde{\boldsymbol{x}}\right)$

Pocket algorithm

- Initialize $\widetilde{\boldsymbol{w}}$ and $\widetilde{\boldsymbol{w}}^{\text{best}} = \mathbf{0}$
- For $t = 1{:}\,T$
  - Loop for all training examples $\widetilde{\boldsymbol{x}}_{\boldsymbol{n}}$ (random order!)
    - Predict $\hat{y}_n = \text{sign}\left(\widetilde{\boldsymbol{w}}^T \widetilde{\boldsymbol{x}}_{\boldsymbol{n}}\right)$
    - If $\hat{y}_n \neq y_n$
      - Update: $\widetilde{\boldsymbol{w}} \leftarrow \widetilde{\boldsymbol{w}} + \eta(y_n \widetilde{\boldsymbol{x}}_{\boldsymbol{n}})$
  - Evaluate $\widetilde{\boldsymbol{w}}$ on the "training data" and calculate the training accuracy
    - If training accuracy by $\widetilde{\boldsymbol{w}}$ is "higher" than the training accuracy by $\widetilde{\boldsymbol{w}}^{\text{best}}$
    - $\widetilde{\boldsymbol{w}}^{\text{best}} \leftarrow \widetilde{\boldsymbol{w}}$
- Output $\widetilde{\boldsymbol{w}}^{\text{best}}$

See lecture 7 for extra details

Gaussian discriminant analysis

GDA

- **Training data:** $D_{tr} = \{\left(\boldsymbol{x}_i \in \mathbb{R}^D,\ \mathrm{y}_i \in \{+1,-1\}\right)\}_{i=1}^{N}$

- **Goal:** construct $p(Y = c|\boldsymbol{x})$ for $\hat{y} = \max\limits_{c\in\{+1,-1\}} p(Y = c|\boldsymbol{x})$

- **Bayes' rules:** $p(Y = c|\boldsymbol{x}) \propto p(\boldsymbol{x}|Y = c)p(Y = c)$
  - $p(Y = c)$: Bernoulli
  - $\mathrm{p}(\boldsymbol{x}|Y = c)$: multi-dimensional Gaussian

Nonlinear GDA

- $p(\boldsymbol{x}|\, Y = +1)$ and $p(\boldsymbol{x}|\, Y = +1)$ have their own covariance matrices $\boldsymbol{\Sigma}_{+1}, \boldsymbol{\Sigma}_{-1}$
- See slides 9, 10 for how to compute them
- See also your homework # 2

Linear GDA

- $p(\boldsymbol{x}|\, Y = +1)$ and $p(\boldsymbol{x}|\, Y = +1)$ share the same covariance matrix $\boldsymbol{\Sigma}$

- Built upon the previous slide, given $\boldsymbol{\Sigma}_{+1}, \boldsymbol{\Sigma}_{-1}$ and let N_{+1}, N_{-1} be the number of training examples per class, $\boldsymbol{\Sigma} = \frac{N_{+1} \times \boldsymbol{\Sigma}_{+1} + N_{-1} \times \boldsymbol{\Sigma}_{-1}}{N}$

- See your homework # 2 for how to compute it

Prediction (please do “log” to prevent overflow)

$$\max_{c \in \{+1,-1\}} p(Y=c|\boldsymbol{x}) = \max_{c \in \{+1,-1\}} p(\boldsymbol{x}|Y=c)p(Y=c)$$

$$= \max_{c \in \{+1,-1\}} \log p(\boldsymbol{x}|Y=c) + \log p(Y=c)$$