6.S093 Visual Recognition through Machine Learning Competition

Joseph Lim and Aditya Khosla

Acknowledgment: Many slides from David Sontag and Machine Learning Group of UT Austin

Image by
kirkh.deviantart.com

What is Machine Learning???

What is Machine Learning???

According to the Wikipedia,
Machine learning concerns the construction and study of systems that can learn from data.

What is ML?

- Classification
  - From data to discrete classes

What is ML?

- Classification
  - From data to discrete classes
- Regression
  - From data to a continuous value

What is ML?

- Classification
  - From data to discrete classes
- Regression
  - From data to a continuous value
- Ranking
  - Comparing items

What is ML?

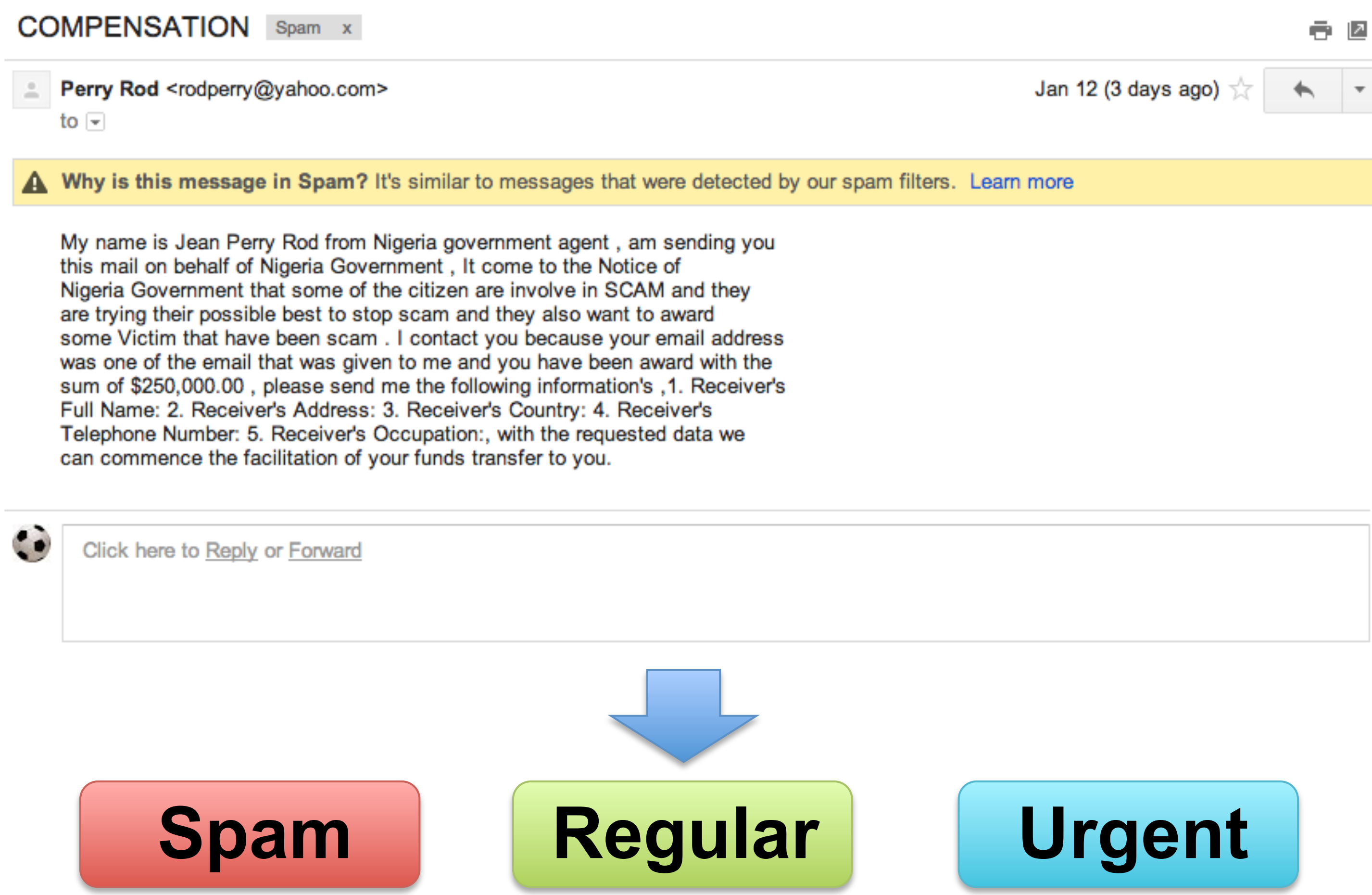
- Classification
  - From data to discrete classes
- Regression
  - From data to a continuous value
- Ranking
  - Comparing items
- Clustering
  - Discovering structure in data

What is ML?

- Classification
  - From data to discrete classes
- Regression
  - From data to a continuous value
- Ranking
  - Comparing items
- Clustering
  - Discovering structure in data
- Others

Classification: data to discrete classes

- Spam filtering



Classification:
data to discrete classes

- Object classification

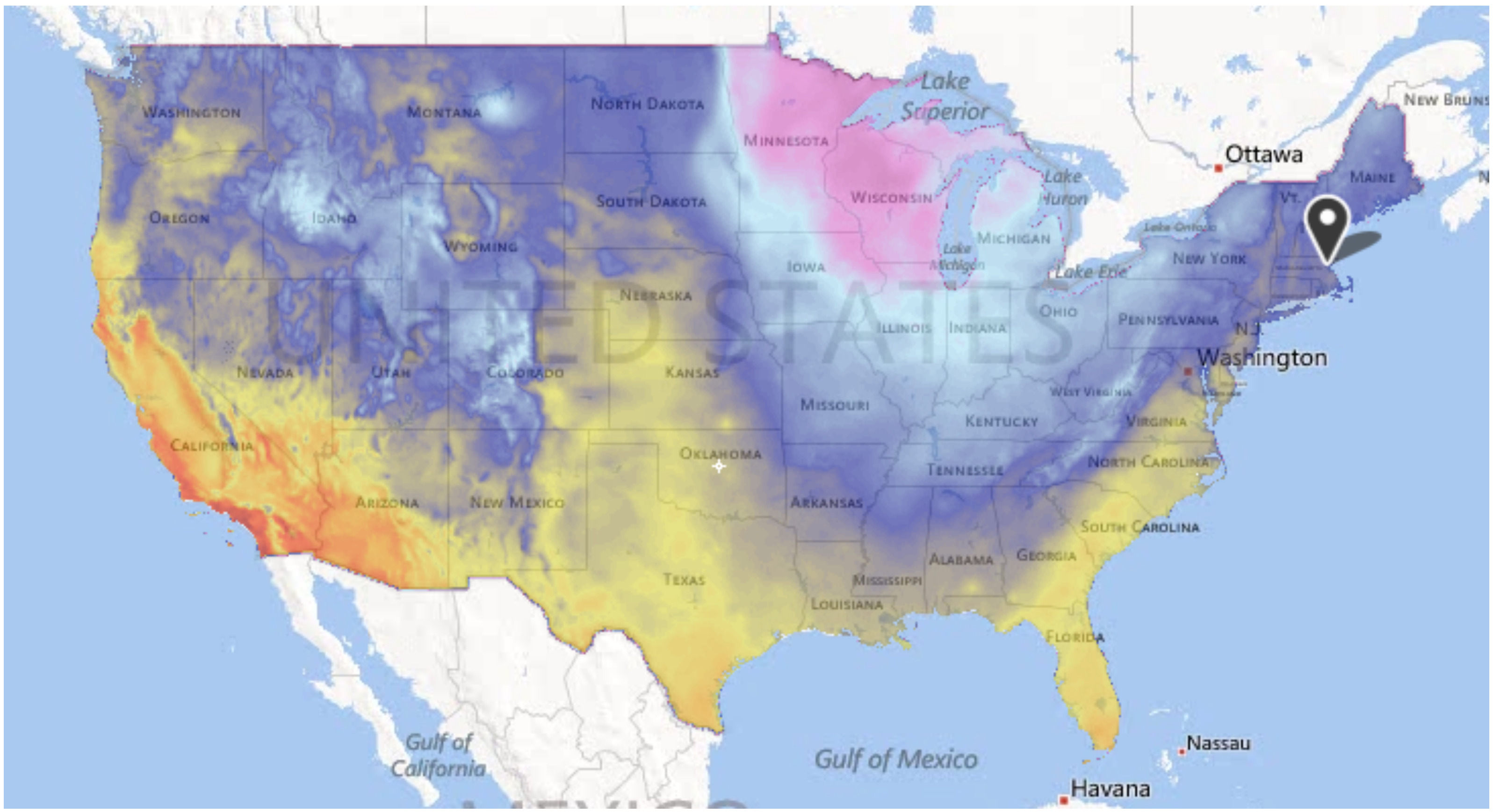
Fries | Hamburger | None

Regression: data to a value

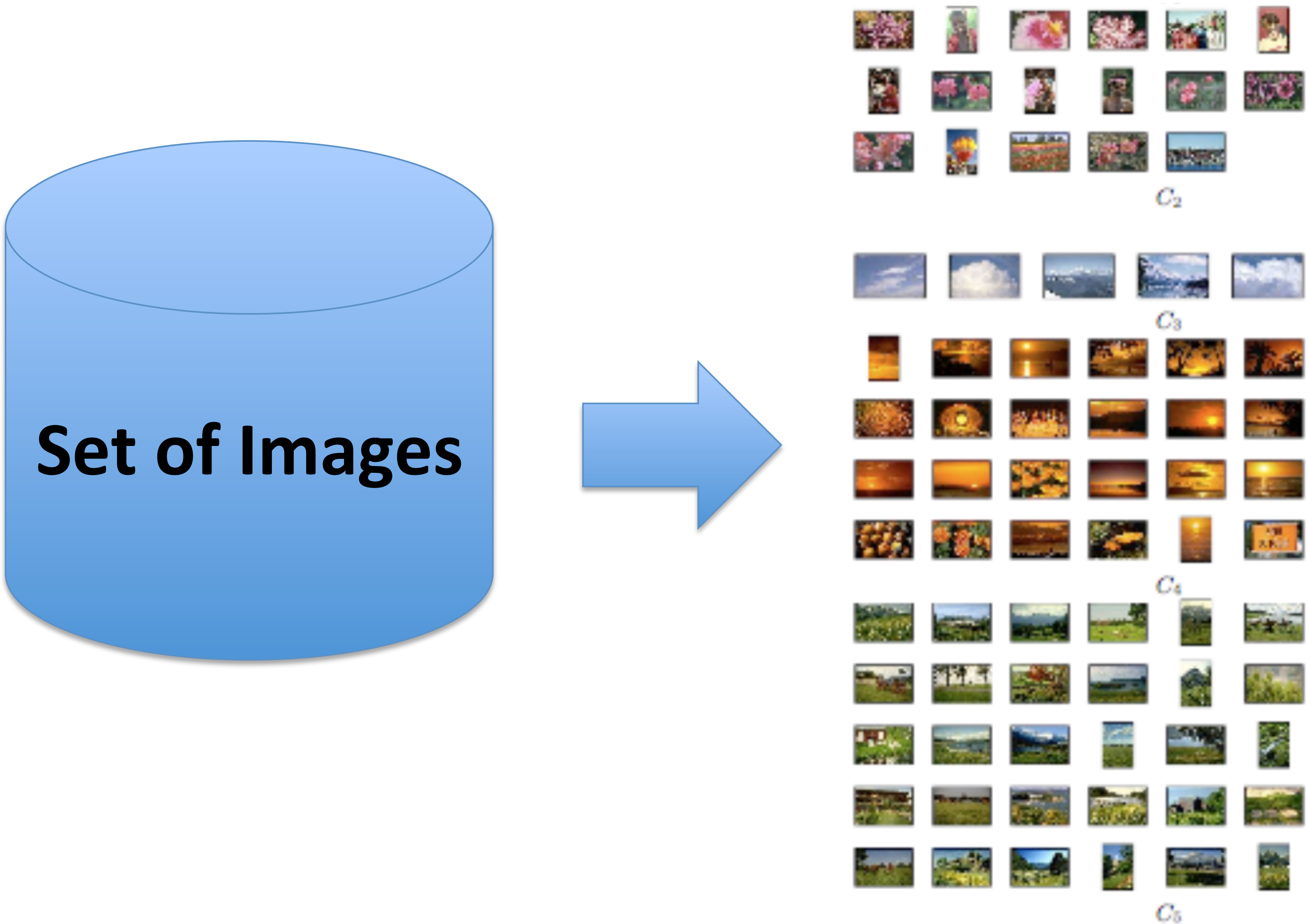
- Stock market

Regression: data to a value

- Weather prediction



Clustering: Discovering structure



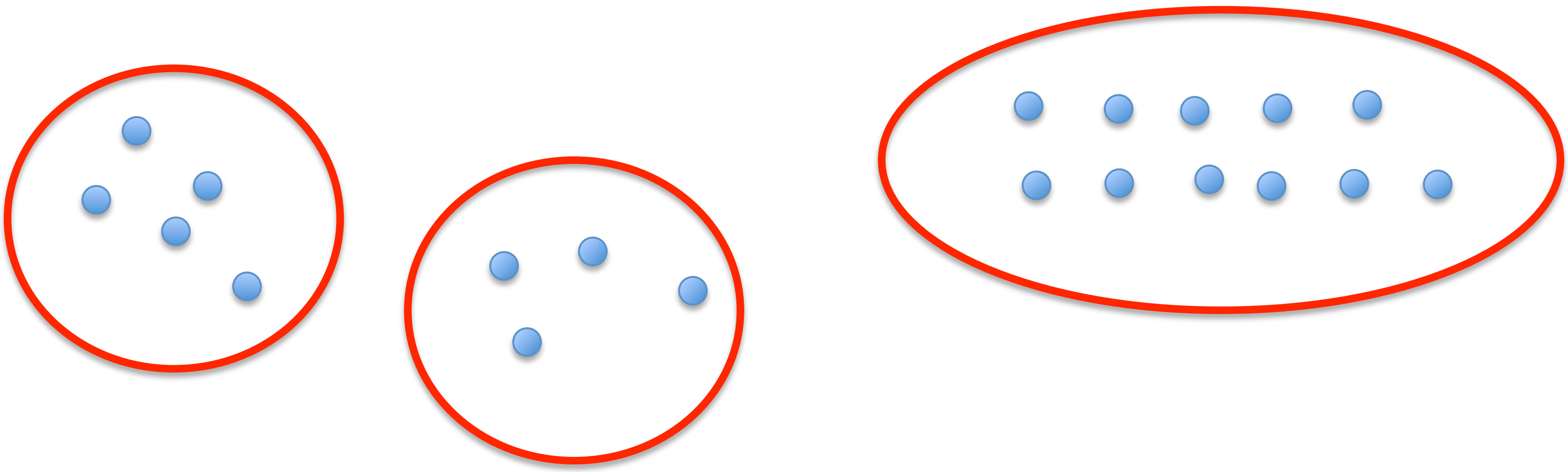
Clustering

Clustering

- Unsupervised learning
- Requires data, but **NO LABELS**
- Detect patterns
  - Group emails or search results
  - Regions of images
- Useful when don’t know what you’re looking for

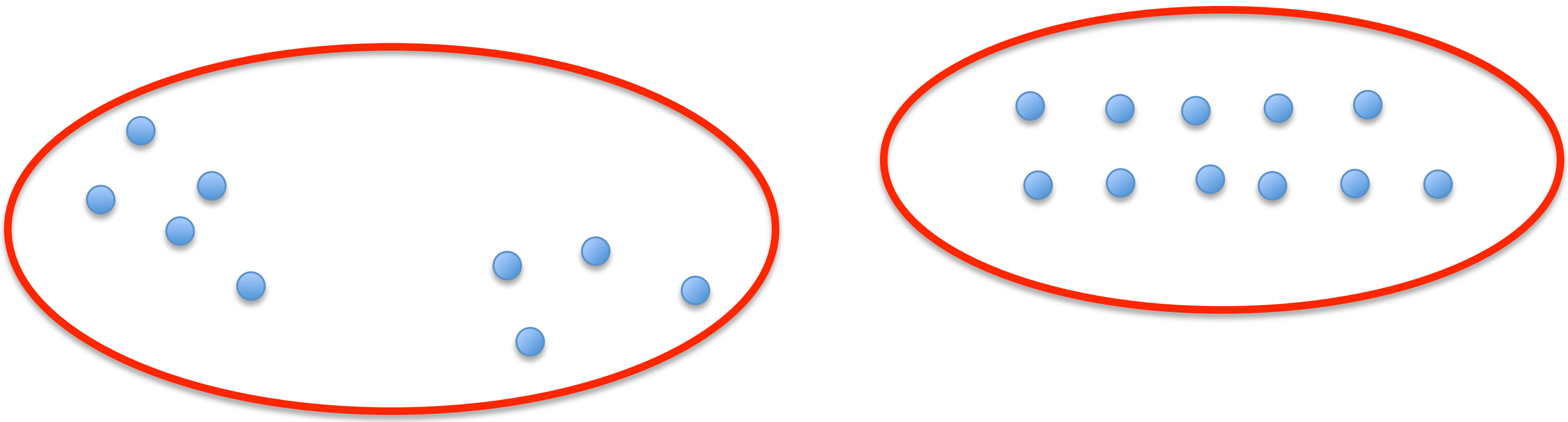
Clustering

- Basic idea: group together similar instances
- Example: 2D point patterns



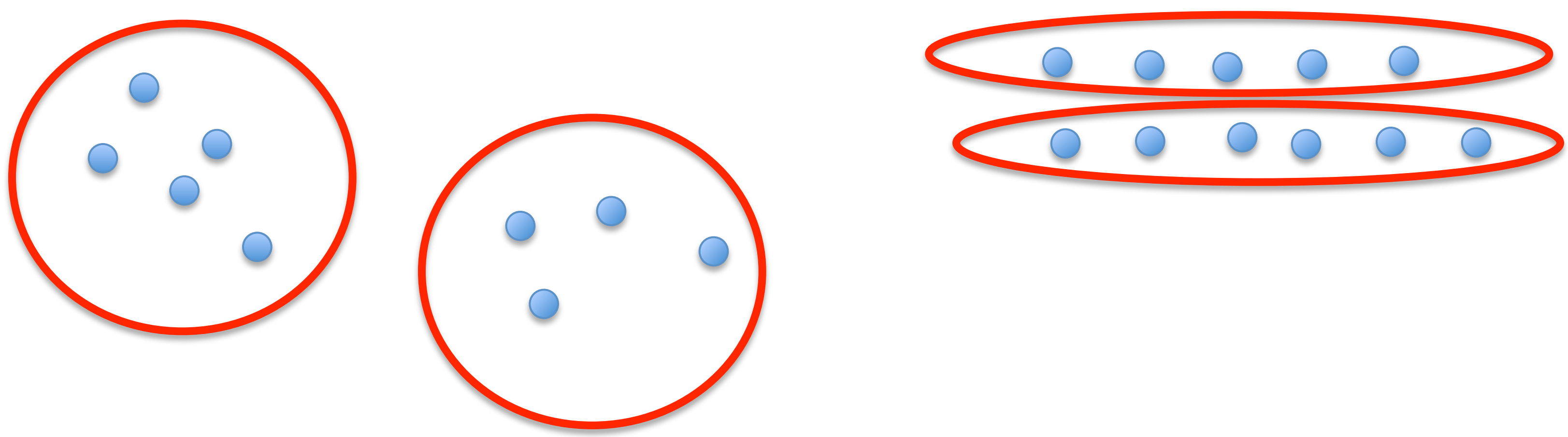
Clustering

- Basic idea: group together similar instances
- Example: 2D point patterns



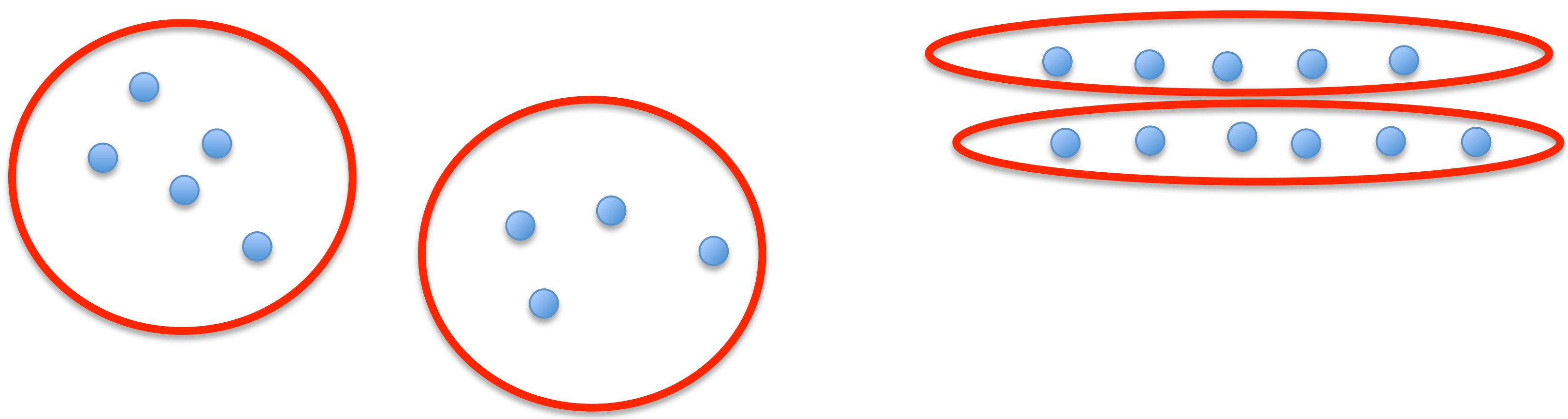
Clustering

- Basic idea: group together similar instances
- Example: 2D point patterns



Clustering

- Basic idea: group together similar instances
- Example: 2D point patterns



- What could "similar" mean?
  - One option: small Euclidean distance
  - Clustering is crucially dependent on the measure of similarity (or distance) between "points"

K-Means

- An iterative clustering algorithm
  - Initialize: Pick K random points as cluster centers
  - Alternate:
    - Assign data points to closest cluster center
    - Change the cluster center to the average of its assigned points
  - Stop when no points' assignments change

Animation is from Andrey A. Shabalin's website

K-Means

- An iterative clustering algorithm
  - Initialize: Pick K random points as cluster centers
  - Alternate:
    - Assign data points to closest cluster center

$$S_i^{(t)} = \{x_p : \|x_p - m_i^{(t)}\|^2 \le \|x_p - m_j^{(t)}\|^2 \ \forall j, 1 \le j \le k\},$$

    - Change the cluster center to the average of its assigned points

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j$$

  - Stop when no points' assignments change

Properties of K-means algorithm

- Guaranteed to converge in a finite number of iterations

- Running time per iteration:
  - Assign data points to closest cluster center

    O(KN)
  - Change the cluster center to the average of its assigned points

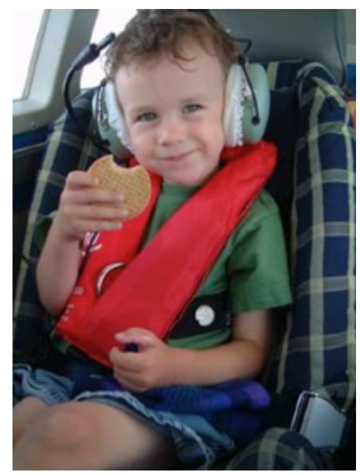
    O(N)

Example: K-Means for Segmentation

K=2

Goal of Segmentation is to partition an image into regions each of which has reasonably homogenous visual appearance.

Original

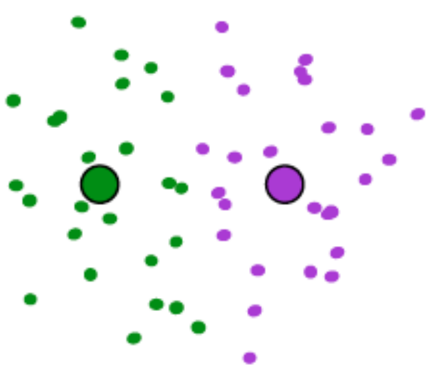


Example: K-Means for Segmentation

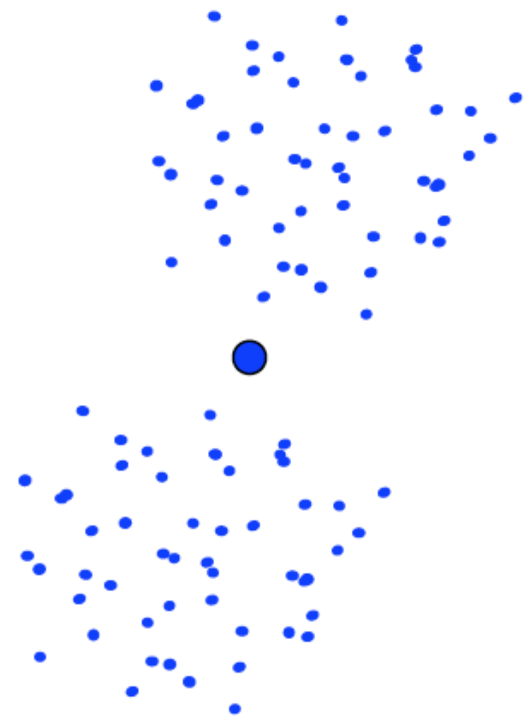
Pitfalls of K-Means

- K-means algorithm is heuristic
  - Requires initial means
  - It does **matter** what you pick!

- K-means can get stuck



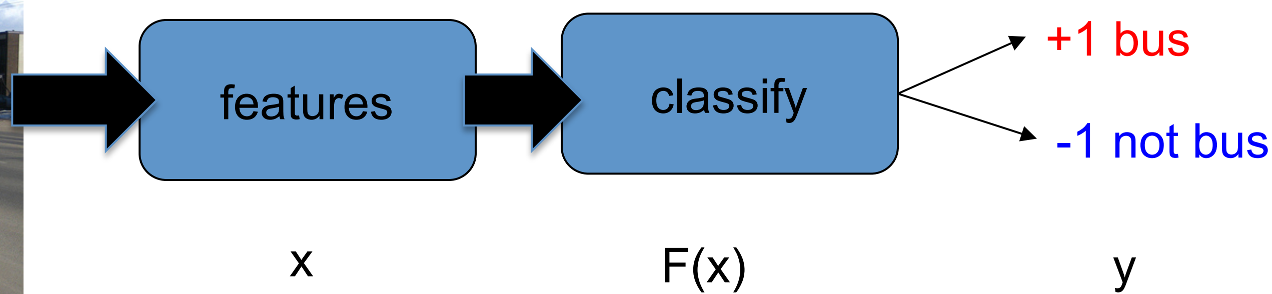
K=1 should be better



K=2 should be better

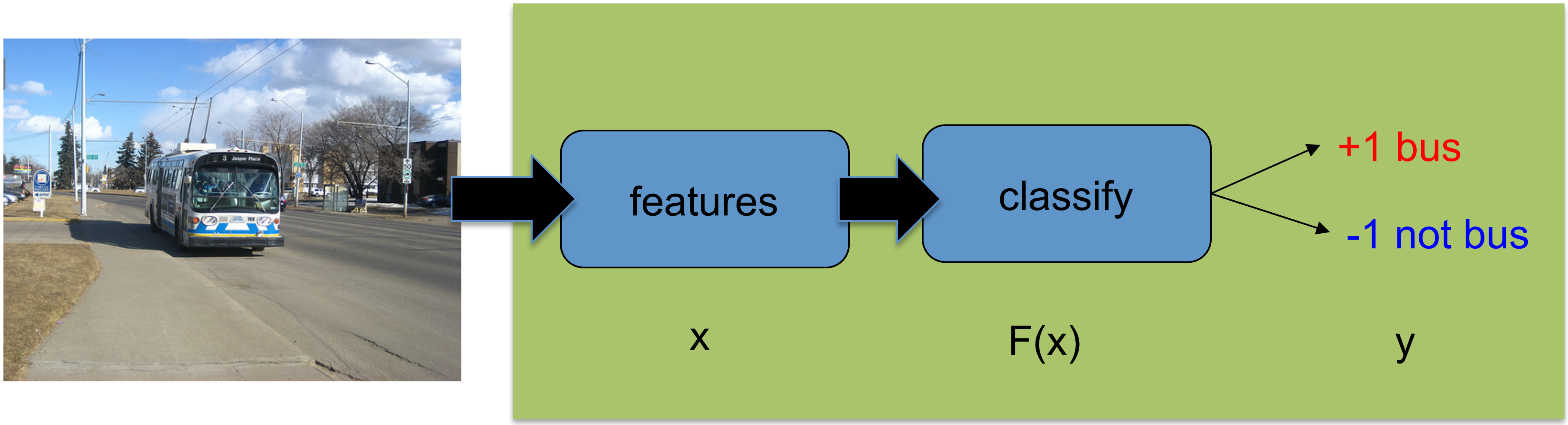
Classification

Typical Recognition System



- Extract features from an image
- A classifier will make a decision based on extracted features

Typical Recognition System



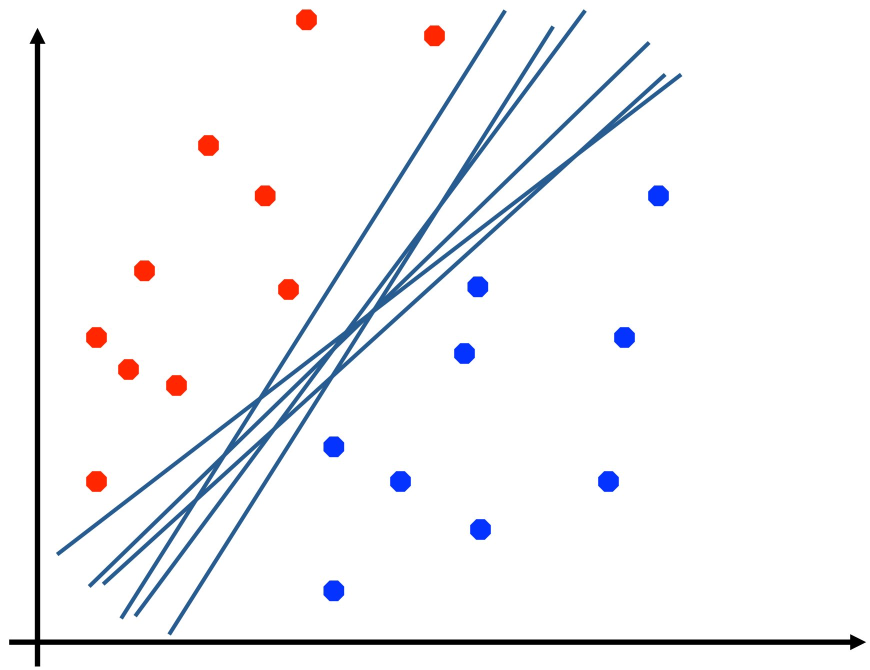
- Extract features from an image
- A classifier will make a decision based on extracted features

Classification

- Supervised learning
- Requires data, AND **LABELS**
- Useful when you know what you're looking for

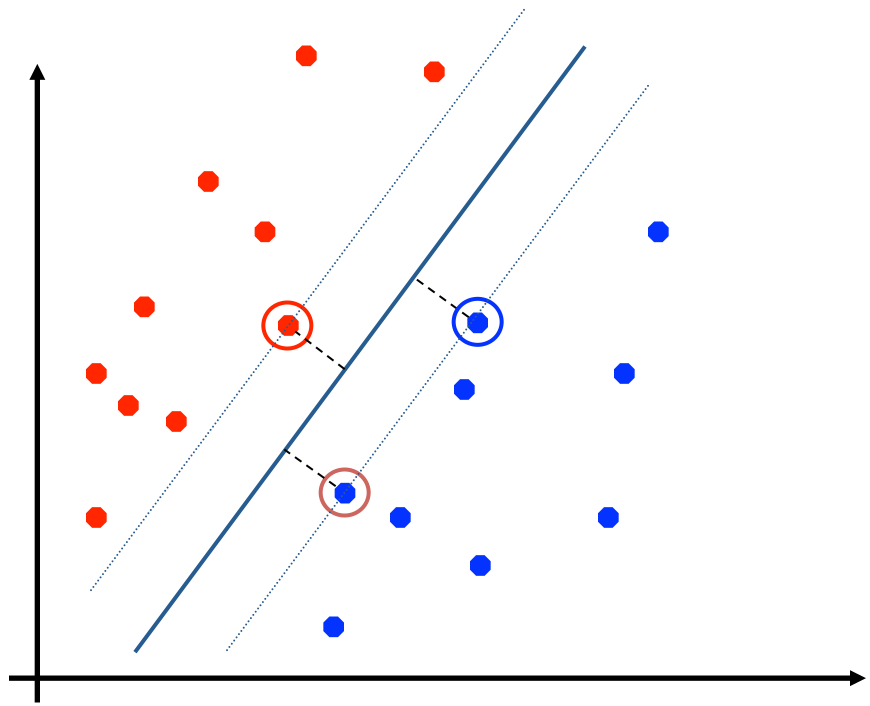
Linear classifier

- Which of these linear separators is optimal?



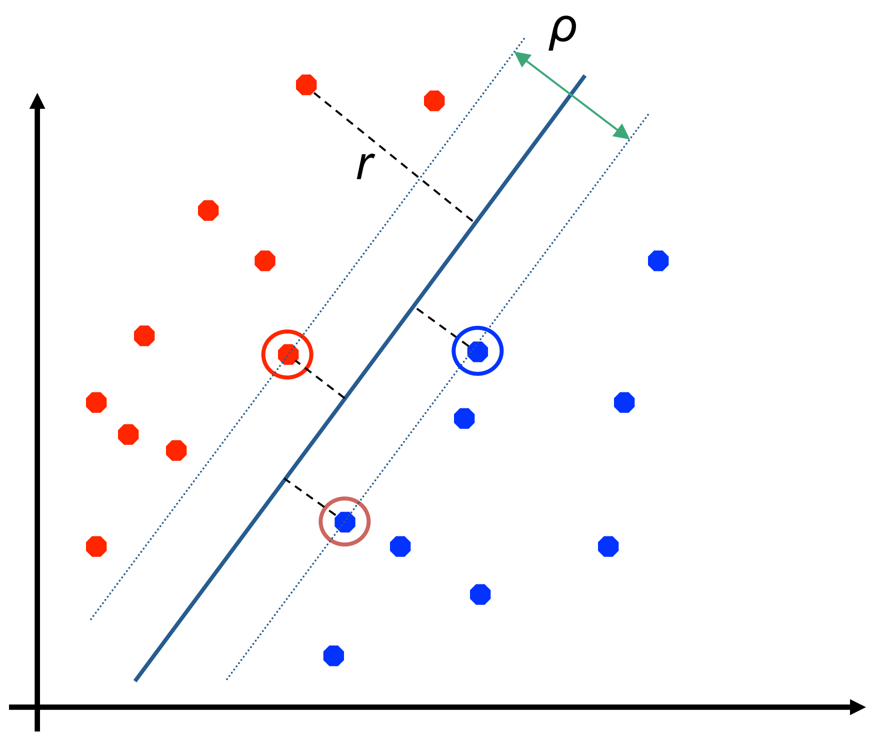
Maximum Margin Classification

- Maximizing the margin is good according to intuition and PAC theory.
- Implies that only support vectors matter; other training examples are ignorable.



Classification Margin

- Distance from example $\mathbf{x}_i$ to the separator is $r = \frac{\mathbf{w}^T\mathbf{x}_i + b}{\|\mathbf{w}\|}$
- Examples closest to the hyperplane are ***support vectors***.
- ***Margin*** ρ of the separator is the distance between support vectors.



Linear SVM Mathematically

- Let training set $\{(\mathbf{x}_i, y_i)\}_{i=1..n}$, $\mathbf{x}_i \in \mathbf{R}^d$, $y_i \in \{-1, 1\}$ be separated by a hyperplane with margin ρ. Then for each training example $(\mathbf{x}_i, y_i)$:

$$\begin{array}{ll} \mathbf{w}^\mathbf{T}\mathbf{x}_i + b \leq -\rho/2 & \text{if } y_i = -1 \\ \mathbf{w}^\mathbf{T}\mathbf{x}_i + b \geq \rho/2 & \text{if } y_i = 1 \end{array} \quad \Leftrightarrow \quad y_i(\mathbf{w}^\mathbf{T}\mathbf{x}_i + b) \geq \rho/2$$

- For every support vector $\mathbf{x}_s$ the above inequality is an equality. After rescaling $\mathbf{w}$ and b by $\rho/2$ in the equality, we obtain that distance between each $\mathbf{x}_s$ and the hyperplane is $r = \frac{y_s(\mathbf{w}^T\mathbf{x}_s + b)}{\|\mathbf{w}\|} = \frac{1}{\|\mathbf{w}\|}$

- Then the margin can be expressed through (rescaled) $\mathbf{w}$ and b as:

$$\rho = 2r = \frac{2}{\|\mathbf{w}\|}$$

Linear SVMs Mathematically (cont.)

- Then we can formulate the *quadratic optimization problem:*

Find **w** and b such that

$$\rho = \frac{2}{\|\mathbf{w}\|} \text{ is maximized}$$

and for all $(\mathbf{x}_i, y_i)$, $i=1..n$: $y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1$

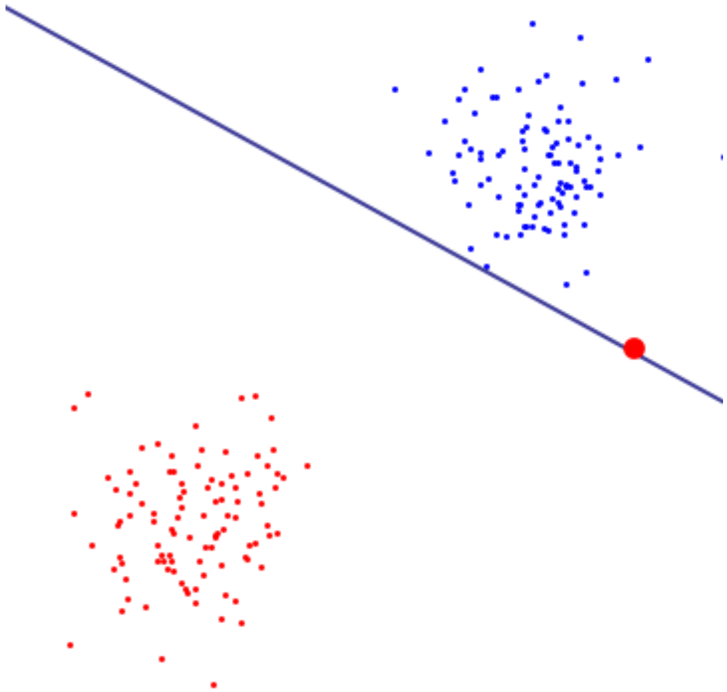
Find **w** and b such that

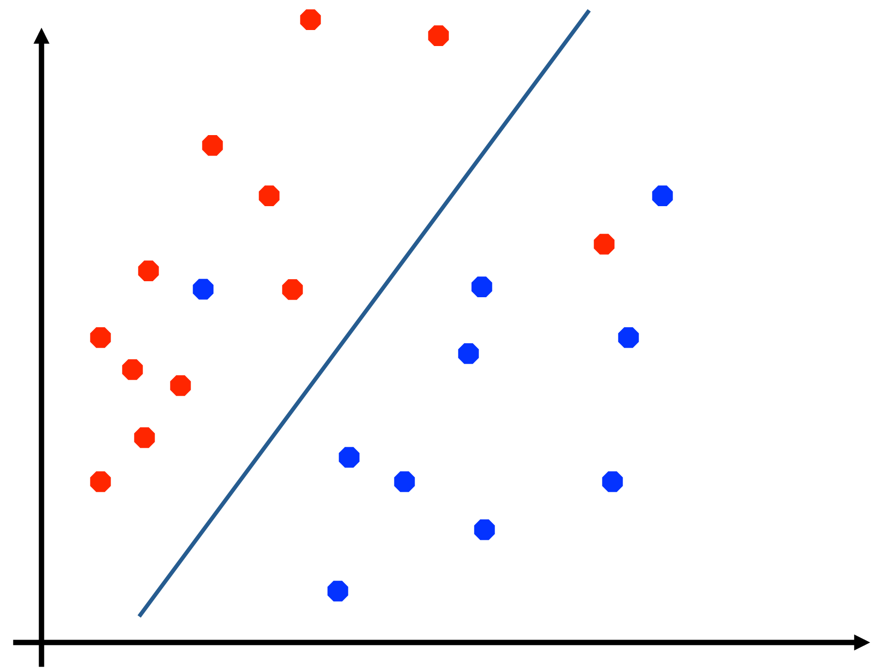
$\mathbf{\Phi}(\mathbf{w}) = ||\mathbf{w}||^2 = \mathbf{w}^T\mathbf{w}$ is minimized

and for all $(\mathbf{x}_i, y_i)$, $i=1..n$: $y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1$

Is this perfect?

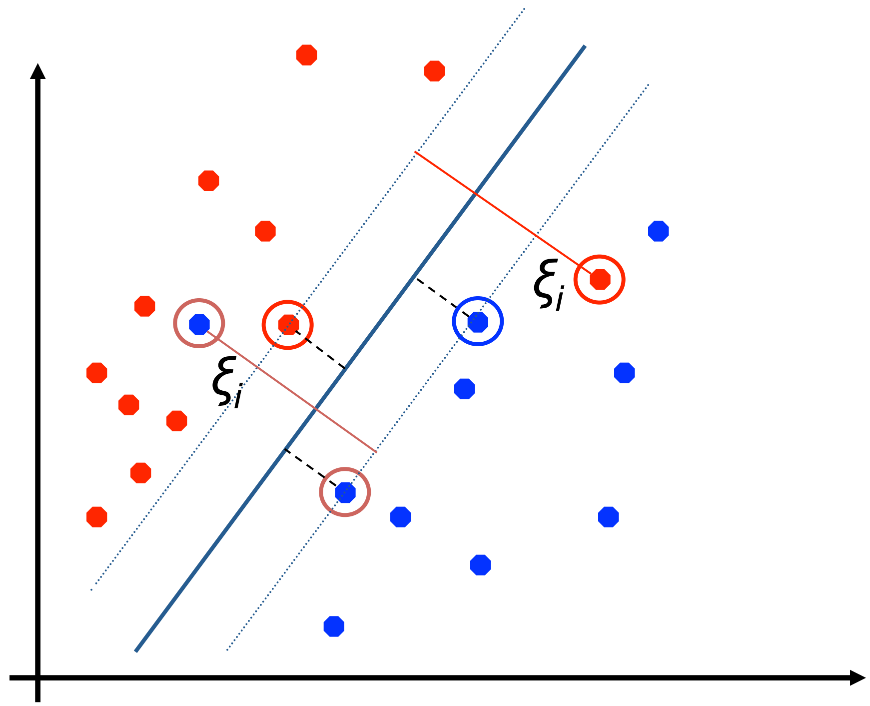
Is this perfect?





Soft Margin Classification

- What if the training set is not linearly separable?
- *Slack variables* ξ_i can be added to allow misclassification of difficult or noisy examples, resulting margin called *soft*.



Soft Margin Classification Mathematically

- The old formulation:

> Find **w** and b such that
> $\mathbf{\Phi}(\mathbf{w}) = \mathbf{w}^T\mathbf{w}$ is minimized
> and for all $(\mathbf{x}_i, y_i)$, $i=1..n$: $y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1$

- Modified formulation incorporates slack variables:

> Find **w** and b such that
> $\mathbf{\Phi}(\mathbf{w}) = \mathbf{w}^T\mathbf{w} + C\Sigma\xi_i$ is minimized
> and for all $(\mathbf{x}_i, y_i)$, $i=1..n$: $y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i$, $\xi_i \geq 0$

- Parameter *C* can be viewed as a way to control overfitting: it "trades off" the relative importance of maximizing the margin and fitting the training data.

Exercise

- Exercise 1. Implement K-means
- Exercise 2. Play with SVM’s C-parameter