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

Set of Images
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
    \[ S_i^{(t)} = \{ x_p : \| x_p - m_i^{(t)} \|^2 \leq \| x_p - m_j^{(t)} \|^2 \ \forall j, 1 \leq j \leq k \}, \]
    • 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

Goal of Segmentation is to partition an image into regions each of which has reasonably homogenous visual appearance.
Example: K-Means for Segmentation

K=2  K=3  K=10  Original
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 \textbf{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 $x_i$ to the separator is $r = \frac{\mathbf{w}^T x_i + b}{\|\mathbf{w}\|}$
- Examples closest to the hyperplane are **support vectors**.
- **Margin** $\rho$ of the separator is the distance between support vectors.
Linear SVM Mathematically

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

  \[
  \begin{align*}
  w^T x_i + b &\leq -\rho/2 \quad \text{if } y_i = -1 \\
  w^T x_i + b &\geq \rho/2 \quad \text{if } y_i = 1 \\
  \iff y_i (w^T x_i + b) &\geq \rho/2 
  \end{align*}
  \]

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

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

  \[ \rho = 2r = \frac{2}{\|w\|} \]
Linear SVMs Mathematically (cont.)

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

Find $\mathbf{w}$ and $b$ such that

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

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 $\mathbf{w}$ and $b$ such that

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

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?
• \textit{Slack variables} $\xi_i$ can be added to allow misclassification of difficult or noisy examples, resulting margin called \textit{soft}.
Soft Margin Classification Mathematically

• The old formulation:
  
  Find \( w \) and \( b \) such that
  
  \[ \Phi(w) = w^T w \] is minimized
  
  and for all \((x_i, y_i), i=1..n:\)
  
  \[ y_i (w^T x_i + b) \geq 1 \]

• Modified formulation incorporates slack variables:
  
  Find \( w \) and \( b \) such that
  
  \[ \Phi(w) = w^T w + C \sum \xi_i \] is minimized
  
  and for all \((x_i, y_i), i=1..n:\)
  
  \[ y_i (w^T x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \]

• 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