Object Recognition by Discriminative Methods

Sinisa Todorovic

1st Sino-USA Summer School in VLPR
July, 2009
Motivation

Car/non-car Classifier

from Kristen Grauman, B. Leibe
Motivation

Discriminate yes, but what?

from Kristen Grauman, B. Leibe
Motivation

Discriminate yes, but what?

Problem 1: What are good features

from Kristen Grauman, B. Leibe
Motivation

from Kristen Grauman, B. Leibe
Motivation

Discriminate yes, but against what?

from Kristen Grauman, B. Leibe
Motivation

Discriminate yes, but against what?

Problem 2: What are good training examples

from Kristen Grauman, B. Leibe
Motivation

from Kristen Grauman, B. Leibe
Motivation

Discriminate yes, but how?

from Kristen Grauman, B. Leibe
Motivation

Discriminate yes, but how?

Problem 3: What is a good classifier

from Kristen Grauman, B. Leibe
Motivation

Sliding windows????

What is a non-car?

Is Bayes decision optimal?

from Kristen Grauman, B. Leibe
How to Classify?

Discriminant function may have a probabilistic interpretation.
How to Classify?

\[ F(x, y) = P(x, y) \]

Generative
How to Classify?

\[ F(x, y) = P(y | x) \]

Discriminative
How to Classify?

observable feature $\xrightarrow{x} F(x, y) \xrightarrow{y} y$ class $\{+1, -1\}$

discriminant function

$F(x, y) = \text{sign}(w.x + b)$

Linear discriminant function
How to Classify?

Large margin classifiers

margin = width that the boundary can be increased before hitting a datapoint

Var 2

Var 1

Large margin classifiers
Given: \( \{(x_1, y_1), \ldots, (x_n, y_n)\} \)
query: \( x \)

\[ y = \hat{y} \quad \text{s.t.} \quad \hat{x} = \arg\min_{x_i} D(x, x_i) \]
Maxim-Margin Linear Classifier

\[ w \cdot x + b = 0 \]

support vectors
Maxim-Margin Linear Classifier

\[ w \cdot x + b = -k \]

\[ w \cdot x + b = 0 \]
Maxim-Margin Linear Classifier

\[ w \cdot x + b = -k \]

\[ w \cdot x + b = k \]

\[ w \cdot x + b = 0 \]

support vectors
Maxim-Margin Linear Classifier

\[ w \cdot x + b = -k \]

\[ w \cdot x + b = k \]

margin width
\[ \frac{2|k|}{\|w\|} \]

support vectors
Maxim-Margin Linear Classifier

Problem:
\[
\max_w \frac{2|k|}{\|w\|} \\
\text{subject to:}
\]
\[
w \cdot x + b \geq k, \ \forall x \text{ of class 1} \\
w \cdot x + b \leq -k, \ \forall x \text{ of class -1}
\]
Maxim-Margin Linear Classifier

Problem:

\[
\max_w \frac{2}{\|w\|}
\]

subject to:

\[
w \cdot x + b \geq 1, \forall x \text{ of class 1}
\]

\[
w \cdot x + b \leq -1, \forall x \text{ of class -1}
\]
\[ w \cdot x + b \geq 1, \ \forall x \text{ of class 1} \]
\[ w \cdot x + b \leq -1, \ \forall x \text{ of class -1} \]

\[ y = \pm 1 \]

\[ y(w \cdot x + b) \geq 1, \ \forall x \]
Problem:
\[
\max_{w} \frac{2}{\|w\|}
\]
\[
s.t.
\]
\[
y(w.x + b) \geq 1, \ \forall x
\]
Problem:

\[
\begin{align*}
\max_w & \quad \frac{2}{\|w\|} \\
\text{s.t.} & \quad y(w \cdot x + b) \geq 1, \quad \forall x
\end{align*}
\]

\[
\downarrow
\]

\[
\begin{align*}
\min_w & \quad \frac{1}{2} \|w\|^2 \\
\text{s.t.} & \quad y(w \cdot x + b) \geq 1, \quad \forall x
\end{align*}
\]
Dual Problem

Solve using Lagrangian

\[ L = \frac{1}{2} w.w - \sum_{i} \alpha_i [y_i(w.x_i + b) - 1] \]
Dual Problem

Solve using Lagrangian

\[ L = \frac{1}{2} w.w - \sum_i \alpha_i [y_i (w.x_i + b) - 1] \]

At solution

\[ \frac{\partial L}{\partial b} = 0 \quad \frac{\partial L}{\partial w} = 0 \]

\[ L = \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i x_j \]
Dual Problem

\[
\max_{\alpha} \sum_{i} \alpha_i - \frac{1}{2} \sum_{i} \sum_{j} \alpha_i \alpha_j y_i y_j x_i x_j
\]

s.t. \[\sum_{i} y_i \alpha_i = 0, \ \alpha_i \geq 0\]
Dual Problem

\[
\max_{\alpha} \sum_{i} \alpha_i - \frac{1}{2} \sum_{i} \sum_{j} \alpha_i \alpha_j y_i y_j x_i x_j \\
\text{s.t.} \quad \sum_{i} y_i \alpha_i = 0, \quad \alpha_i \geq 0
\]
Dual Problem

\[
\max_{\alpha} \sum_{i} \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j x_i x_j
\]

s.t. \[
\sum_i y_i \alpha_i = 0, \quad \alpha_i \geq 0
\]

Then compute:

\[\hat{\alpha}_i\]

\[\hat{\alpha}_i > 0 \quad \text{Only for support vectors}\]

\[\hat{b} = y_i - \hat{w} x_i \quad \text{for support vectors}\]
Linearly Non-Separable Case

Trade-off between maximum separation and misclassification

\[
\min_{\mathbf{w}} \frac{1}{2} \| \mathbf{w} \|^2 + C \sum \gamma_i
\]

s.t.

\[y_i (\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \gamma_i, \quad \forall \mathbf{x}_i\]

\[\gamma_i \geq 0\]
Non-Linear SVMs

- Non-linear separation by mapping data to another space
- In SVM formulation, data appear only in the vector product
- No need to compute the vector product in the new space
- Mercer kernels

\[ x_i x_j \rightarrow K(x_i, x_j) = \Phi(x_i)\Phi(x_j) \]
Vision Applications

- Pedestrian detection
  - multiscale scanning windows
  - for each window compute the wavelet transform
  - classify the window using SVM

“A general framework for object detection,”
C. Papageorgiou, M. Oren and T. Poggio -- CVPR 98
“A general framework for object detection,”
C. Papageorgiou, M. Oren and T. Poggio -- CVPR 98
Shortcomings of SVMs

- Kernelized SVM requires evaluating the kernel for a test vector and each of the support vectors

- Complexity = Kernel complexity $\times$ number of support vectors

- For a class of kernels this can be done more efficiently
  
  $[\text{Maji, Berg, Malik CVPR 08}]$

\[
K(x_i, x_j) = \sum_{n} \min(x_i(n), x_j(n))
\]

intersection kernel
Outline

Random Forests

Nearest Neighbor

Breiman, Fua, Criminisi, Cipolla, Shotton, Lempitsky, Zisserman, Bosch, ...

Shakhnarovich, Viola, Darrell, Berg, Frome, Malik, Todorovic...
Distance Based Classifiers

Given: \( \{(x_1, y_1), \ldots, (x_n, y_n)\} \)

query: \( x \)

\[
y = \hat{y} \quad \text{s.t.} \quad \hat{x} = \arg\min_{x_i} D(x, x_i)
\]
Given query $\mathcal{X}$ and datapoint-class pairs $\{(x_1, y_1), \ldots, (x_n, y_n)\}$

Learn a Mahalanobis distance metric that

- brings points from the same class closer, and
- makes points from different classes be far away

“Distance metric learning with application to clustering with side information”
E. Xing, A. Ng, and M. Jordan, NIPS, 2003.
\[
\min_A \sum_{(x_i, x_j) \in S} \|x_i - x_j\|_A
\]
\[
\text{s.t. }
\sum_{(x_i, x_j) \in D} \|x_i - x_j\|_A \geq 1
\]
Learning Global Distance Metric

$$\min_A \sum_{(x_i, x_j) \in S} \| x_i - x_j \|_A$$

s.t.

$$\sum_{(x_i, x_j) \in D} \| x_i - x_j \|_A \geq 1$$

PROBLEM!
Learning Global Distance Metric

Problem with multimodal classes

before learning

after learning
Per-Exemplar Distance Learning

BEFORE

local neighborhood

AFTER

margin

margin

Similarly labeled
Differently labeled
Differently labeled

Frome & Malik ICCV07, Todorovic & Ahuja CVPR08
Distance Between Two Images

\[ D(F, I) = \sum_{j} w^F_j d^F_j (I) = w^F \cdot d^F (I) \]

distance between j-th patch in image F and image I
Learning from Triplets

For each image $I$ in the set we have:

$$w^F \cdot d^F (I^D) > w^F \cdot d^F (I^S)$$
For each image $I$ in the set we have:

$$w^F \cdot d^F (I^D) > w^F \cdot d^F (I^S)$$

$$\downarrow$$

$$x_i = d^F (I^D) - d^F (I^S)$$
For each image $I$ in the set we have:

$$w^F \cdot d^F(I^D) > w^F \cdot d^F(I^S)$$

$$\downarrow$$

$$x_i = d^F(I^D) - d^F(I^S)$$

$$\downarrow$$

$$w^F \cdot x_i > 0$$
Max-Margin Formulation

- Learn for each focal image $F$ independently

$$\arg \min_{w^F, \gamma} \frac{1}{2} \| w^F \|^2 + C \sum_i \gamma_i$$

s.t.

$$\forall i, \ w^F \cdot x_i \geq 1 - \gamma_i$$

$$w^F \geq 0 \quad \gamma_i \geq 0$$
Max-Margin Formulation

• Learn for each focal image $F$ independently

\[
\arg\min_{w^F, \gamma} \frac{1}{2} \|w^F\|^2 + C \sum_i \gamma_i
\]

s.t.
\[
\forall i, \quad w^F \cdot x_i \geq 1 - \gamma_i
\]
\[
w^F \geq 0 \quad \gamma_i \geq 0
\]

PROBLEM! \forall i \rightarrow \text{only some images}

They heuristically select 15 closest images
Max-Margin Formulation

\[ D(F, I) = \sum_j w^F_j d^F_j(I) = w^F \cdot d^F(I) \]

distance between j-th patch in image F and image I

**Problem:**
After computing \( w^F \) in-class and out-of-class datapoints that have initially been closest may not be closest after learning.
EM-based Max-Margin Formulation of Local Distance Learning

discovery of subcategories in segmentation trees

axis = subcategory
trees = points in the feature space of subcategories

linear classifier
rescaled axes

Two categories of image points

Zero relevance
Max relevance

Todorovic & Ahuja CVPR08
Learning from Triplets

For each image I in the set:

Frome, Malik ICCV07:

\[ w^F \cdot d^F (I^D) > w^F \cdot d^F (I^S) \]

Todorovic et al. CVPR08:

\[ w^F \cdot d^F (I) Pr (I \in D) > w^F \cdot d^F (I) Pr (I \in S) \]
Published, best categorizations on Caltech-101 and Caltech-256

Caltech-101:
- Ours
- Bosch-ICCV07
- Frome-ICCV07
- Zhang-CVPR06
- Lazebnik-CVPR06
- Todorovic-CVPR06

Caltech-256:
- Ours
- Bosch-ICCV07
- CaltechTechRep07
- Todorovic-CVPR06
Decision Trees – Not Stable

- Partitions of data via recursive splitting on a single feature
- Result: Histograms based on data-dependent partitioning
- Majority voting
  - Quinlan C4.5; Breiman, Freedman, Olshen, Stone (1984); Devroye, Gyorfi, Lugosi (1996)
RF = Set of decision trees such that each tree depends on a random vector sampled independently and with the same distribution for all trees in RF
Hough Forests

Combine:

spatial info + class info

“Class-Specific Hough Forests for Object Detection”
Juergen Gall and Victor Lempitsky
CVPR 2009
Hough Forests

Binary Tests Selection

- Test with optimal split:
  \[\arg\min_k \left( U_\ast(\{p_i \mid t^k(\mathcal{I}_i) = 0\}) + U_\ast(\{p_i \mid t^k(\mathcal{I}_i) = 1\}) \right)\]

- Class-label uncertainty:
  \[U_1(A) = |A| \cdot \text{Entropy}(\{c_i\})\]

- Offset uncertainty:
  \[U_2(A) = \sum_{i: c_i = 1} (d_i - d_A)^2\]

- Interleaved: Type of uncertainty is randomly selected for each node
In the test image all features cast votes about the location of the bounding box

“Class-Specific Hough Forests for Object Detection”
Juergen Gall and Victor Lempitsky
CVPR 2009
Hough Forests

Pedestrians (TUD)

“Class-Specific Hough Forests for Object Detection”
Juergen Gall and Victor Lempitsky
CVPR 2009
Thank you!