Boosting with Incomplete Information

Gholamreza Haffari[1] Yang Wang[1] Shaojun Wang[2]
Greg Mori[1] Feng Jiao[3]

[1]Simon Fraser University, Canada
[2]Wright State University, USA
[3]Yahoo! Inc., USA

The 25th International Conference on Machine Learning
Helsinki, Finland, 2008

Introduction

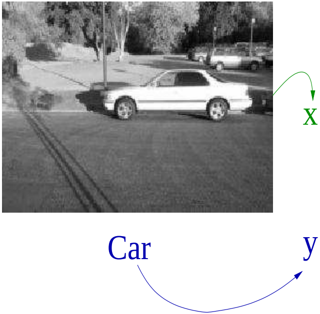
Supervised Classification

Given data set $\mathcal{D} = \{x_i, y_i\}$, x_i is the input vector, y_i is the class label, learn a mapping function $\mathcal{F} : \mathcal{X} \to \mathcal{Y}$

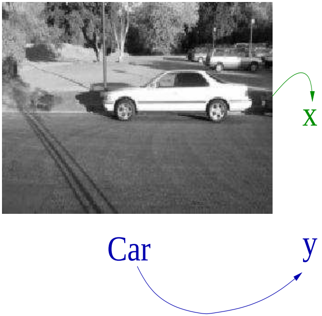
Classification with Incomplete Information

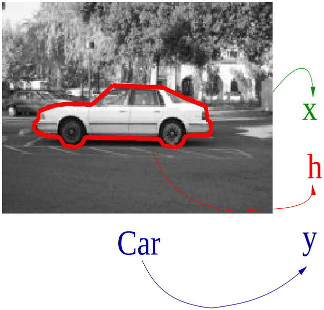
- Given two kinds of data sets $\mathcal{D}_1 = \{x_i, y_i\}$, $\mathcal{D}_2 = \{x_j, h_j, y_j\}$, learn a mapping function $\mathcal{F} : \mathcal{X} \times \mathcal{H} \to \mathcal{Y}$
- This two data sets assumption is general and can be applied to many problems.

Motivation



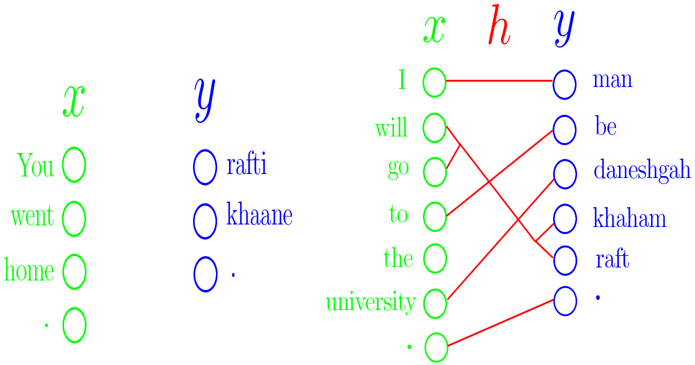
Motivation





Motivation

Motivation


x
y
You
went
home
.
rafti
khaane
.
x
h
y
I
will
go
to
the
university
.
man
be
daneshgah
khaham
raft
.

Previous Work

- EM algorithm for generative models
- Max margin classification (Bi $\&$ Zhang, 2004;Chechik et al., 2007)
- Hidden conditional random fields (Koo $\&$ Collins, 2005; Quattoni et al., 2005)
- Second order cone programming (Shivaswamy et al., 2006)

Review of boosting

Basics

- Feature(weak learner,sufficent statistics): $f_k : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$
- Final classifier: $y^* = \arg\max_y \left(\sum_k \lambda_k f_k(x,y) \right)$

Learning parameters λ_k

- Unnormalized model
  - Minimize $\sum_{x_i} \sum_y q_\lambda(y|x_i)$
  - where $q_\lambda(y|x) := \exp \sum_k \lambda_k \left[f_k(x,y) - f_k(x,\tilde{y}_x) \right]$
- Normalized model
  - Maximize $\sum_{x_i} \log p_\lambda(\tilde{y}_{x_i}|x_i)$
  - where $p_\lambda(y|x) := q_\lambda(y|x)/Z_\lambda(x)$

(Lebanon & Lafferty, 2002)

Primal/Dual Problem

Definition

- (extended) KL divergence:
  $D(p,q) := \sum_x \tilde{p}(x) \sum_y \left(p(y|x) \log \frac{p(y|x)}{q(y|x)} - p(y|x) + q(y|x) \right)$
- feasible set:
  $\mathcal{F}(\tilde{p}, f) = \left\{ p | \sum_x \tilde{p}(x) \sum_y p(y|x)(f_j(x,y) - E_{\tilde{p}}[f_j|x]) = 0, \forall j \right\}$

Primal problems

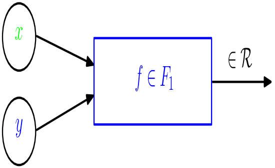
$$\begin{array}{lll} \text{(P1) min.} & D(p, q_0) & \\ \text{s.t.} & p \in \mathcal{F}(\tilde{p}, f) & \end{array} \qquad \begin{array}{ll} \text{(P2) min.} & D(p, q_0) \\ \text{s.t.} & p \in \mathcal{F}(\tilde{p}, f) \\ & \sum_y p(y|x) = 1 \ \forall x \end{array}$$

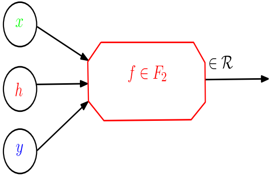
(Lebanon $\&$ Lafferty, 2002)

Problem Statement

- Data sets: $\mathcal{D}_1 = \{(x_i, y_i)\}$, $\mathcal{D}_2 = \{(x_j, h_j, y_j)\}$, $|\mathcal{D}_1| >> |\mathcal{D}_2|$ in general
- Features:

$$\mathcal{F}_1 = \{f_k(x, y)\} \qquad \mathcal{F}_2 = \{f_k(x, h, y)\}$$





- Goal: how to learn a classifer using $\mathcal{D}_1 \cup \mathcal{D}_2$ and $\mathcal{F}_1 \cup \mathcal{F}_2$?

Boosting with Hidden Variables

Normalized model

- Model: $p_\lambda(y|x,h) \propto e^{\boldsymbol{\lambda}_1^T \cdot [\mathbf{f}_1(x,y) - \mathbf{f}_1(x,\tilde{y}_x)] + \boldsymbol{\lambda}_2^T \cdot [\mathbf{f}_2(x,h,y) - \mathbf{f}_2(x,h,\tilde{y}_x)]}$
- Objective: maximize the log-likelihood

$$\mathcal{L}(\lambda) := \sum_i \log p_\lambda(y_i|x_i) + \gamma \sum_j \log p_\lambda(y_j|x_j, h_j)$$

Unnormalized model

- Model:$q_\lambda(y|x,h) := e^{\boldsymbol{\lambda}_1^T \cdot [\mathbf{f}_1(x,y) - \mathbf{f}_1(x,\tilde{y}_x)] + \boldsymbol{\lambda}_2^T \cdot [\mathbf{f}_2(x,h,y) - \mathbf{f}_2(x,h,\tilde{y}_x)]}$
- Objective: minimize the exponential loss

$$\mathcal{E}(\lambda) := \sum_i \sum_h q_0(h|x) \sum_y q_\lambda(y|x_i, h) + \gamma \sum_j \sum_y q_\lambda(y|x_j, h_j)$$

Primal/Dual Programs

Definitions

- extended KL-divergence
  $KL(\mathbf{p}||\mathbf{r}) =$
  $\sum_{x,h} \tilde{p}(x) q_0(h|x) \sum_y p(y|h,x) \Big[\log \frac{p(y|x,h)}{r(x,h,y)} - 1 \Big] + r(x,h,y)$
- feasible set $\mathcal{S}(\tilde{\mathbf{p}}, \mathbf{q}_0, \mathcal{F}) = \Big\{ \mathbf{p} \in$
  $\mathcal{M} \Big| \sum_x \tilde{p}(x) \mathbb{E}_{q_0(h|x)p(y|x,h)} \Big[f - \mathbb{E}_{\tilde{p}(y|x)}[f] \Big] = 0, \forall f \in \mathcal{F} \Big\}$

Primal problems

(P1) min. $KL(\mathbf{p}||\mathbf{r})$
s.t. $\mathbf{p} \in \mathcal{S}$

(P2) min. $KL(\mathbf{p}||\mathbf{r})$
s.t. $\mathbf{p} \in \mathcal{S}$
$\sum_y p(y|x,h) = 1 \;\; \forall x, h$

Learning and Inference

Learning

- Construct auxillary function to bound the change of $\mathcal{E}(\lambda + \Delta\lambda) - \mathcal{E}(\lambda)$ or $\mathcal{L}(\lambda) - \mathcal{L}(\lambda + \Delta\lambda)$
- Both parallel and sequential update rules can be derived

Inference

- If h is observed on test data, $y^* = \arg\max p(y|h, x)$
- If h is unobserved on test data, $y^* = \arg\max p(y|x)$. This requires summing over h.

Experiments: Visual Object Recognition

airplane car face motorbike

Experiments: Visual Object Recognition

- 1000 training/testing images, 4 categories
- 30% fully observed training images
- Baselines algorithms

BL1

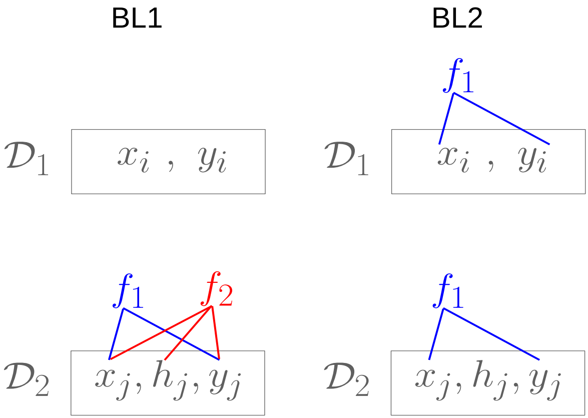
$\mathcal{D}_1$ $\boxed{x_i \;,\; y_i}$

f_1 f_2

$\mathcal{D}_2$ $\boxed{x_j, h_j, y_j}$

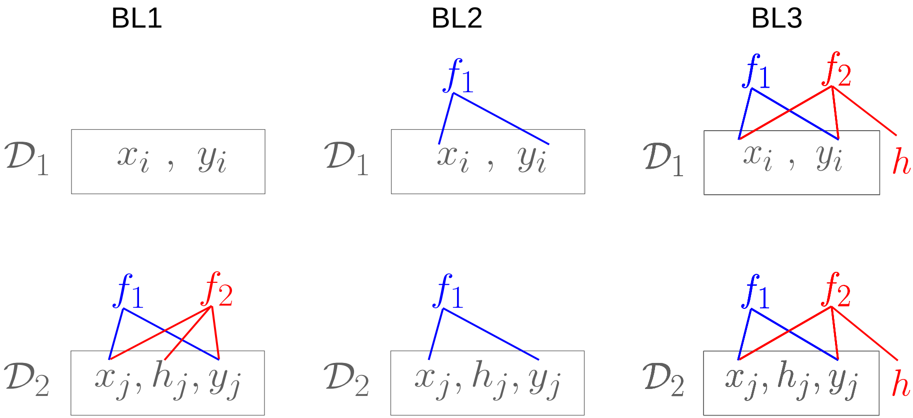
Experiments: Visual Object Recognition

- 1000 training/testing images, 4 categories
- 30% fully observed training images
- Baselines algorithms

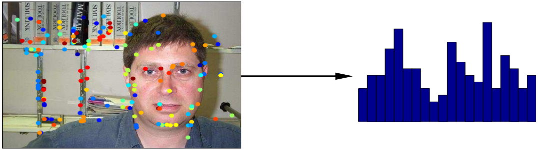


Experiments: Visual Object Recognition

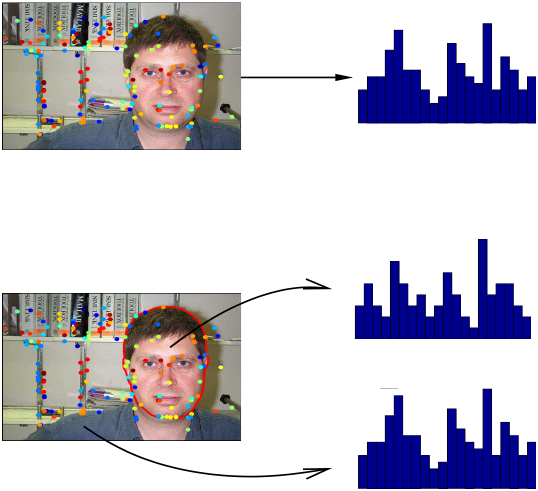
- 1000 training/testing images, 4 categories
- 30% fully observed training images
- Baselines algorithms



Experiments: Visual Object Recognition



Experiments: Visual Object Recognition



Experiments: Visual Object Recognition

	accuracy	log-likelihood
Our method	97.22%	-0.0916
BL1	89.26%	-1.1417
BL2	88.01%	-0.5698
BL3	90.43%	-0.4375

normalized model

	accuracy	log of loss
Our method	94.83%	-0.7412
BL1	82.57%	-1.1231
BL2	89.86%	-0.7977
BL3	87.64%	-0.8068

unnormalized model

Experiments: Named Entity Recognition

- CoNLL03 shared task: 5000 fully observed, 6000 partially observed, 1000 testing
- Features:
  - Lexical: word forms and their positions in the window
  - Syntactic: part-of-speech tags(if available)
  - Orthographic: capitalized, include digits,...
  - Affixes: suffixes and prefixes
  - Left predict: predicted labels for the two previous words

h is unobserved on test data

	f-measure	log-likelihood
Our method	49.45%	-0.5784
BL1	46.63%	-0.5932
BL2	48.10%	-0.5803
BL3	47.80%	-0.5880

normalized model

	f-measure	log of loss
Our method	49.04%	-2.6337
BL1	46.24%	-2.6458
BL2	47.58%	-2.6378
BL3	46.39%	-2.6434

unnormalized model

h is observed on test data

	f-measure	log-likelihood
Our method	59.60%	-0.5759
BL1	56.51%	-0.5916

normalized model

	f-measure	log of loss
Our method	60.17%	-0.2586
BL1	55.46%	-0.2655

unnormalized model

Summary

Conclusion

A boosting approach that extends the traditional boosting framework by incorporating hidden variables, and achieves better results than baseline approaches.

Future work

- Extension to more complex dependent hidden variables (e.g., trees, graphs), variational methos (e.g., loopy BP) may be used
- Connection with confidence-rated AdaBoost (Schapire & Singer,1999)