Univariate Margin Tree

Olcay Taner Yıldız

Department of Computer Engineering
Işık University
Istanbul, Turkey

ISCIS 2010

Outline

1. **Motivation**

2. **Univariate Margin Tree**

3. **Experiments**

4. **Conclusions**

Decision Trees

- Tree-based structures
- Each internal node implements a decision function, $f_m(\boldsymbol{x})$
- Univariate decision tree, $f_m(\boldsymbol{x})$ uses only one attribute
- If attribute is continuous, there will be 2 children of each internal node

Maximum Margin Classifiers

- Become popular in recent years
- Concept of *margin*: The smallest distance between the decision boundary and the closest data points (support vectors)
- Model parameters determined by setting up a convex optimization problem
- Use hinge loss instead of misclassification error
- 2-class problems, extended to *K*-class via one-vs-one, one-vs-rest, etc.

Previous Work on Univariate Decision Trees

- C4.5 tries to minimize the impurity or maximize the information gain
- LDT assumes the two class groups are normally distributed and tries to maximize the ratio of between class distance to within class distance
- Proposal: Univariate Margin Tree express finding best split as a minimization problem.

Setting

- N labeled points $S = ((\boldsymbol{x}^1, y^1), \ldots, (\boldsymbol{x}^N, y^N)) \in (X \times Y)^N$
- X is the input space, a continuous vectorial space of dimension d.
- Y is the label set, $\{-1, +1\}$.
- The data pairs $(\boldsymbol{x}^i, y^i)$ are i.i.d.

Multivariate Margin vs. Univariate Margin

$$
\begin{aligned}
Min \quad & \tfrac{1}{2}||\mathbf{w}||^2 + C\sum_t \epsilon^t \\
s.t. \quad & y^t(\mathbf{w}^T\mathbf{x}^t + w_0) \geq 1 - \epsilon^t \\
& \epsilon^t \geq 0
\end{aligned}
$$

$$
\begin{aligned}
Min \quad & \sum_t \epsilon^t \\
s.t. \quad & y^t(x_j^t + w_0) \geq C - \epsilon^t \\
& \epsilon^t \geq 0
\end{aligned}
$$

- Quadratic programming problem
- ϵ^t and $\mathbf{w}$ are variables.
- C, penalty factor

- Linear programming problem
- ϵ^t and w_0 are variables.
- $|C|$, length of the margin

Linear Programming

- Neither primal nor dual easily solvable
- Search all possible split points w_0 exhaustively
- The inequality $y^t(x_j^t + w_0) \geq C - \epsilon^t$ can be written as

$$\epsilon^t \geq C - x_j^t y^t - w_0 y^t$$

- Setting the left side of the equation to zero ($\epsilon^t = 0$) gives us all possible values for $w_0 = \dfrac{C - x_j^t y^t}{y^t}$

- Calculate other ϵ^j's and check for the maximum value of the objective function $\displaystyle\sum_t \epsilon^t$

The Algorithm

```
Split UnivariateMarginTreeBestSplit(N, d, S, V)
1   bestError = +∞
2   for C = -2.0 to 2.0 step 0.1
3       for i = 1 to d
4           minepsilon = +∞
5           for j = 1 to N
6               w_0 = (C - x_i^j y^j) / y^j
7               sumepsilon = 0.0
8               for k = 1 to N
9                   ε^k = C - x_i^k y^k - w_0 y^k
10                  if ε^k > 0
11                      sumepsilon += ε^k
12              if sumepsilon < minepsilon
13                  bestw_0 = w_0
14          error = ErrorOfSplit(x_i + bestw_0 ≥ 0, V)
15          if error < bestError
16              bestError = error
17              bestSplit = x_i + bestw_0 ≥ 0
18  return bestSplit
```

Datasets

Dataset	d	K	N	Source	Dataset	d	K	N	Source
ads	1558	2	3279	UCI	braintumor1	5920	5	90	Bio
breast	9	2	699	UCI	braintumor2	10367	4	50	Bio
bupa	6	2	345	UCI	dermatology	34	6	366	UCI
dlbcl	5469	2	77	Bio	ecoli	7	8	336	UCI
german	24	2	1000	UCI	glass	9	6	214	UCI
haberman	3	2	306	UCI	iris	4	3	150	UCI
heart	13	2	270	UCI	letter	16	26	20000	UCI
hepatitis	19	2	155	UCI	leukemia1	5327	3	72	Bio
ironosphere	34	2	351	UCI	leukemia2	11225	3	72	Bio
magic	10	2	19020	UCI	lungtumor	12600	5	203	Bio
musk2	166	2	6598	UCI	ocr	256	10	600	UCI
parkinsons	22	2	195	UCI	optdigits	64	10	3823	UCI
pima	8	2	768	UCI	pageblock	10	5	5473	UCI
polyadenylation	169	2	6371	UCI	pendigits	16	10	7494	UCI
prostatetumor	10509	2	102	Bio	segment	19	7	2310	UCI
ringnorm	20	2	7400	UCI	shuttle	9	7	58000	UCI
satellite47	36	2	2134	UCI	srbct	2308	4	83	Bio
spambase	57	2	4601	UCI	vehicle	18	4	846	UCI
transfusion	4	2	748	UCI	wave	21	3	5000	UCI
twonorm	20	2	7400	UCI	winequality	11	7	6497	UCI
9tumors	5726	9	60	Bio	wine	13	3	178	UCI
11tumors	12533	11	174	Bio	yeast	8	10	1484	UCI
14tumors	15009	26	308	Bio	zipcodes	256	10	7291	UCI
					zoo	16	7	101	UCI

Error Rate (2-Class)

Dataset	C4.5	LDT	UMT
ads	**3.36±0.39**	3.65±0.42	3.90±0.30
breast	6.92±1.24	6.45±0.51	**6.32±1.45**
bupa	**38.53±4.38**	40.78±4.42	41.98±5.19
dlbcl	23.33±8.91	25.56±5.37	**21.48±7.16**
german	29.43±1.18	29.28±1.40	**28.20±1.56**
haberman	26.37±0.31	27.25±2.84	**26.18±0.93**
heart	30.89±3.98	**29.00±4.76**	32.44±2.81
hepatitis	22.31±3.97	21.73±1.30	**19.62±2.53**
ironosphere	14.19±4.44	14.87±4.51	**11.28±5.85**
magic	**17.13±0.40**	17.70±0.28	19.21±0.41
musk2	**4.70±0.60**	5.29±0.60	12.70±1.24
parkinsons	15.38±4.47	15.54±5.40	**14.92±3.63**
pima	28.87±2.72	28.44±4.25	**26.26±1.29**
polyadenylation	30.58±2.11	**29.59±1.73**	29.65±1.27
prostatetumor	**15.71±2.43**	20.29±10.64	22.00±11.43
ringnorm	**12.04±0.71**	23.13±1.18	15.12±2.02
satellite47	**14.59±1.33**	15.21±1.17	14.61±0.76
spambase	9.32±1.22	**9.19±0.80**	10.14±0.56
transfusion	24.00±0.00	**23.76±0.76**	24.00±0.00
twonorm	**17.50±0.64**	17.73±0.69	19.14±0.68

Tree Complexity (2-Class)

Dataset	C4.5	LDT	UMT
ads	38.80±19.81	50.20±21.60	**31.00±11.83**
breast	11.50±7.65	10.00±5.66	**7.60±3.41**
bupa	18.70±15.13	13.30±16.34	**8.20±6.96**
dlbcl	3.40±1.90	2.50±2.12	**1.90±1.45**
german	**4.30±6.99**	10.90±20.88	9.10±8.72
haberman	4.60±8.69	4.60±8.69	**2.20±3.79**
heart	9.40±5.44	8.20±5.14	**4.90±2.85**
hepatitis	10.90±9.70	**2.50±3.24**	4.90±4.48
ironosphere	**11.80±6.66**	12.40±9.03	12.10±5.84
magic	86.80±28.47	121.30±44.42	**63.40±13.33**
musk2	114.10±25.18	131.50±28.43	**24.40±11.12**
parkinsons	9.10±5.84	8.80±4.05	**5.80±3.52**
pima	15.10±13.27	23.80±24.83	**7.60±5.06**
polyadenylation	**38.50±20.11**	54.70±37.35	81.70±28.44
prostatetumor	5.20±2.10	4.90±2.02	**3.70±0.95**
ringnorm	157.00±44.70	261.10±67.42	**70.90±10.96**
satellite47	40.60±16.42	27.40±20.29	**26.50±9.72**
spambase	79.90±31.78	106.60±38.93	**54.70±17.69**
transfusion	**1.00±0.00**	1.90±2.85	**1.00±0.00**
twonorm	225.10±44.95	248.80±51.27	**171.70±23.60**

Error Rate (*K*-Class)

Dataset	C4.5	LDT	UMT(OVO)	UMT(OVR)
9tumors	85.45±1.92	84.55±4.89	78.64±10.51	**74.09±8.58**
11tumors	46.56±6.05	41.80±5.42	40.98±2.04	**34.10±4.69**
14tumors	77.39±5.75	75.32±3.85	70.54±2.44	**63.42±4.40**
braintumor1	37.50±1.47	**33.44±5.32**	38.75±5.15	39.69±9.55
braintumor2	51.11±8.20	53.89±12.02	**44.44±10.14**	53.33±12.88
dermatology	12.48±1.52	12.48±1.52	11.12±1.38	**9.76±2.26**
ecoli	21.30±2.70	18.70±3.36	23.04±3.51	**16.52±2.56**
glass	37.43±3.43	**36.35±4.79**	45.27±3.50	40.95±2.78
iris	7.65±1.11	7.45±0.83	**7.06±1.89**	7.06±2.95
letter	18.44±0.50	19.26±0.57	23.14±0.81	**13.76±0.80**
leukemia1	15.60±2.27	**14.80±3.79**	20.40±12.57	15.20±4.92
leukemia2	**14.40±5.06**	17.60±3.37	30.00±14.88	16.00±7.54
lungtumor	15.65±6.03	**13.33±8.53**	20.72±5.55	14.49±4.73
ocr	23.65±4.18	23.65±4.18	22.50±2.76	**19.80±3.85**
optdigits	14.99±0.69	14.57±1.48	11.71±0.49	**8.50±0.77**
pageblock	3.72±0.48	**3.63±0.45**	4.79±0.51	4.50±0.36
pendigits	5.90±0.47	6.32±0.58	9.78±0.73	**4.66±0.45**
segment	**4.97±1.14**	5.44±0.67	8.43±1.27	6.32±0.98
shuttle	**0.07±0.01**	0.11±0.02	1.07±2.24	0.39±0.07
srbct	29.31±9.92	23.10±11.84	26.21±8.93	**14.83±7.46**
vehicle	29.82±2.43	30.35±2.34	32.15±3.43	**28.80±1.71**
wave	24.67±1.17	24.14±0.58	24.52±0.80	**23.19±1.49**
winequality	45.58±1.26	46.40±0.76	46.12±0.86	**45.10±0.76**
wine	10.33±3.58	13.00±6.89	9.17±2.97	**8.50±3.09**
yeast	47.85±3.16	49.50±2.20	45.06±2.23	**42.71±0.94**
zipcodes	15.60±0.82	17.30±0.84	12.27±0.61	**10.02±0.85**
zoo	**15.95±4.50**	**15.95±4.50**	18.65±5.62	18.11±3.83

Summary

- Proposed a novel univariate decision tree classifier
- At each decision node set up a convex optimization problem
- No need for quadratic optimization, linear optimization enough
- For a given C, search model parameters exhaustively
- Apply the usual univariate decision tree algorithm recursively

Conclusions

- For two class problems, UMT performs as good as C4.5 and LDT in terms of accuracy
- For two class problems, UMT produces significantly smaller trees than C4.5 and LDT
- For *K*-class problems, one-vs-rest UMT is significantly better than C4.5 and LDT