Seminar “Maschinellem Lernen”

An Improved Model Selection Heuristic for AUC

Tutor: Jan-Nikolas Sulzmann
Jiawei Du

Overview

- Evaluate Scoring Classifiers
- ROC & AUC
- sROC & sAUC
- Experimental Results
- Conclusions

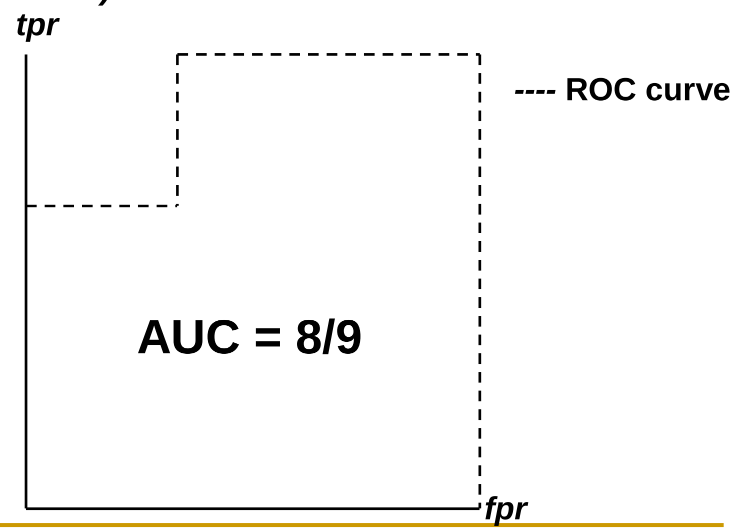
Evaluate Scoring Classifiers

- Classification Models ≈ Class Decision or Score
  - *Classification performance*
    - *Accuracy* $= \dfrac{tp + tn}{P + N}$
  - *Probability estimation performance*
    - *Brier score* $= \sum_x (p'(x) - p(x))^2$
  - *Ranking performance*
    - *ROC curve & AUC*
      - Include all possible thresholds
      - estimates probability that randomly chosen positive example is ranked before randomly chosen negative example

ROC & AUC

1. Calculate score for each instance in the dataset
2. Rank instances on decreasing score
3. Draw ROC curve
   1. next instance is + : move 1/P up
   2. next instance is – : move 1/N to the right
4. Calculate the area under the curve (AUC)

instance	score	
1	1.0	+
2	0.9	+
3	0.6	–
4	0.5	+
5	0.2	–
6	0.0	–



Calculate AUC without ROC

- Calculate AUC directly from the sorted test instances, without the need for drawing an ROC curve or calculating ranks

- P positive instances
- N negative instances
- $\{y_1, \ldots, y_P\}$ is score for the positive instances
- $\{x_1, \ldots, x_N\}$ is score for the negative instances

- AUC counts the number of pairs of positives and negatives such that the former has higher score than the latter

Calculate AUC without ROC

- Ψ_{ij} is 1 if $y_i - x_j > 0$, and 0 otherwise

$$AUC = \frac{1}{PN} \sum_{i=1}^{P} \sum_{j=1}^{N} \Psi_{ij}$$

- Z be the sequence produced by sorting $\{y_1,\ldots,y_P\} \cup \{x_1,\ldots,x_N\}$ in descending order

$$AUC = \frac{1}{PN} \sum_{j=1}^{N} (s_j - j) = \frac{1}{PN} \sum_{j=1}^{N} \sum_{t=1}^{s_j - j} 1$$

- s_j is the rank of x_j in Z , and $s_j - j$ is the number of positives before the j th negative in Z , namely the number of positives correctly ranked relative to each negative

An Example (AUC)

Classifier M1

instance	score	
1	1.0	+
2	0.7	+
3	0.6	+
4	0.5	–
5	0.4	–
6	0.0	–

$$\boxed{AUC = \frac{1}{3*3}(3+3+3) = \frac{9}{9}}$$

Classifier M2

instance	score	
1	1.0	+
2	0.9	+
3	0.6	–
4	0.5	+
5	0.2	–
6	0.0	–

$$AUC = \frac{1}{3*3}(2+3+3) = \frac{8}{9}$$

M1 gets the highest AUC

AUC Deteriorate

- subtract 0.25 from the positive scores

Classifier M1

instance	score	
1	0.75	+
2	0.45	+
3	0.35	+
4	0.5	–
5	0.4	–
6	0.0	–

instance	score	
1	0.75	+
4	0.5	–
2	0.45	+
5	0.4	–
3	0.35	+
6	0.0	–

Classifier M2

instance	score	
1	0.75	+
2	0.65	+
3	0.6	–
4	0.25	+
5	0.2	–
6	0.0	–

$$AUC = \frac{1}{3*3}(1+2+3) = \frac{6}{9}$$

$$\boxed{AUC = \frac{8}{9}}$$

- AUC deteriorates when positive scores are decreased

sROC & sAUC

- sROC curve study the relationship between AUC and differences in the score values

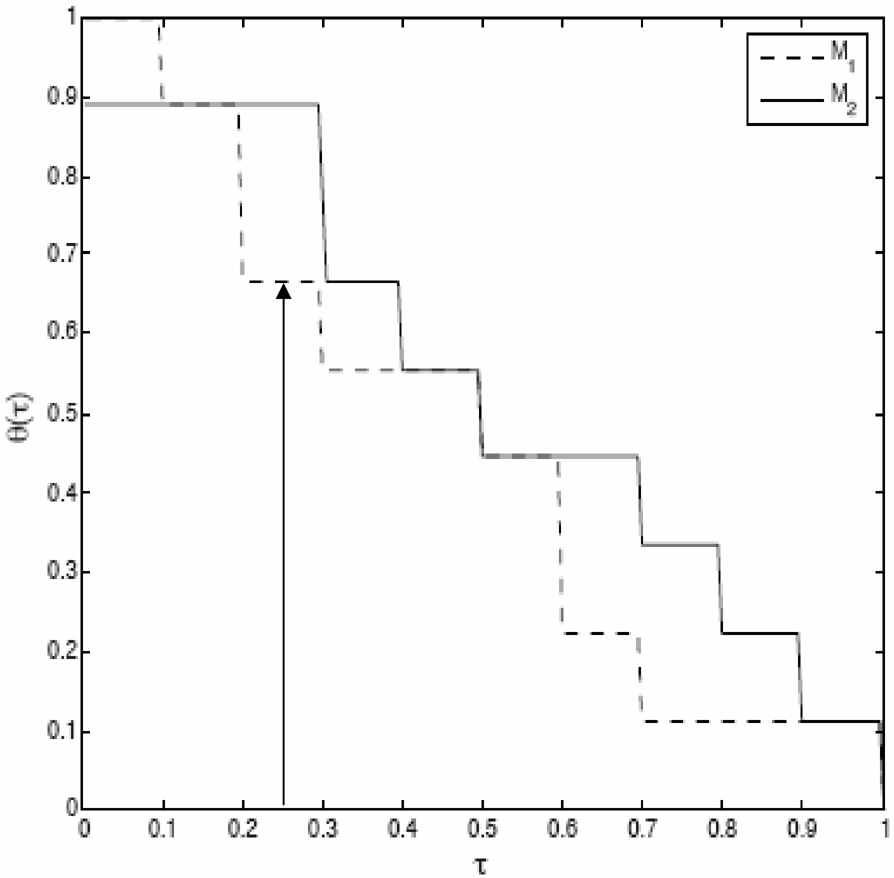
$$AUC = \frac{1}{PN} \sum_{i=1}^{P} \sum_{j=1}^{N} \Psi_{ij}(\tau)$$

- ~~Ψ_{ij} is 1 if $y_i - x_j > 0$, and 0 otherwise~~
- Ψ_{ij} is 1 if $y_i - x_j > \tau$, and 0 otherwise
- AUC counts the number of pairs of positives and negatives such that the former has higher score (at least τ)than the latter

Compare Classifiers in sROC Curve

τ =0 AUC(M1)=1
AUC(M2)=0.89

τ =0.25 AUC(M1)=0.67
AUC(M2)=0.89



Calculate sAUC without sROC

- sAUC is a measure of how rapidly the AUC deteriorates with increasing margin τ

$$sAUC = \int_0^1 \frac{1}{PN} \sum_{i=1}^{P} \sum_{j=1}^{N} \Psi_{ij}(\tau) d\tau$$

$$= \frac{1}{PN} \left(\sum_{i=1}^{P} \sum_{t=1}^{r_i - i} y_i - \sum_{j=1}^{N} \sum_{t=1}^{s_j - j} x_j \right)$$

- The number of negative instances that correctly ranked relative to each positive instance * the score of this positive instance
- The number of positive instances that correctly ranked relative to each negative instance * the score of this negative instance

An Example (sAUC)

Classifier M1

instance	score	
1	1.0	+
2	0.7	+
3	0.6	+
4	0.5	–
5	0.4	–
6	0.0	–

$$sAUC = \frac{1}{3*3}(3*1.0 + 3*0.7 + 3*0.6) - \frac{1}{3*3}(3*0.5 + 3*0.4 + 3*0.0) \approx 0.47$$

Classifier M2

instance	score	
1	1.0	+
2	0.9	+
3	0.6	–
4	0.5	+
5	0.2	–
6	0.0	–

$$\boxed{sAUC = \frac{1}{3*3}(3*1.0 + 3*0.9 + 2*0.5) - \frac{1}{3*3}(2*0.6 + 3*0.2 + 3*0.0) \approx 0.54}$$

M2 is robust over a larger range of margins

Difference

- ROC Curve & AUC
  - ordinal comparison between the scores
  - only ranking information
  - overfitting the validation data

- sROC Curve & sAUC
  - not only ranking information
  - but also score information
    (the difference between the scores)

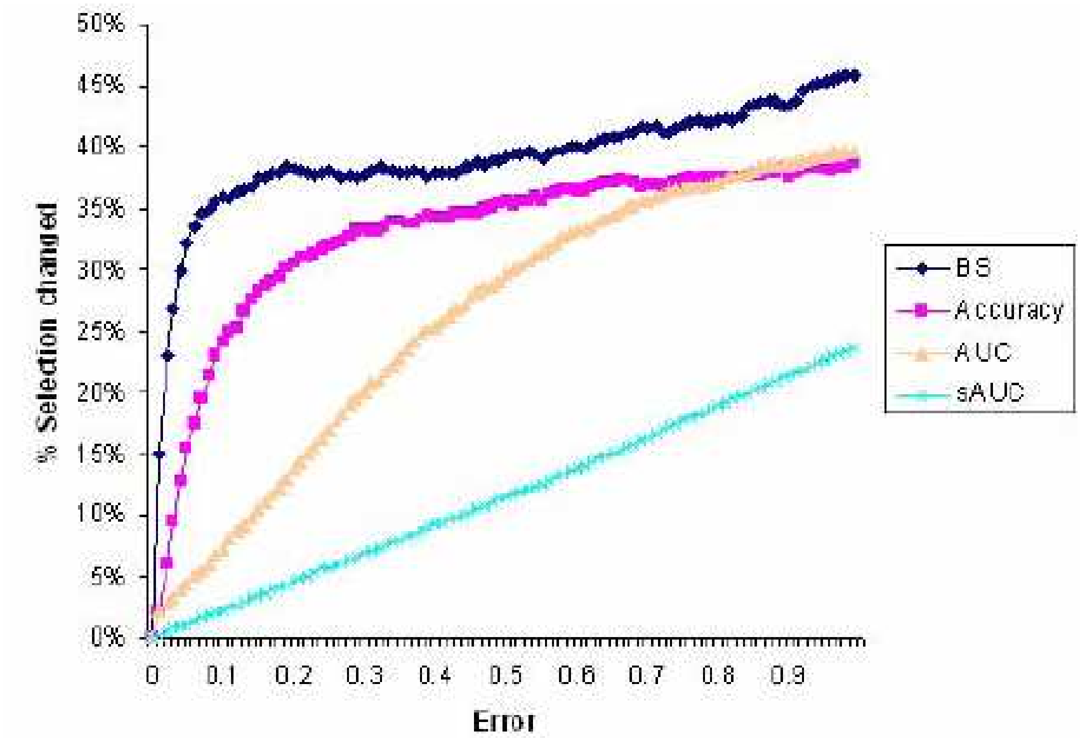
Experiment 1

- Goal: sAUC outperforms AUC and Brier score for selecting models, particularly when validation data is limited

- Two experiments:
  - *artificial data*
    1. Data set A & B (100 instances)
    2. Each instance gets a probability p in [0,1]
    3. Label instance (+ if p >= 0.5)
    4. Swap 10 examples of data set *A*, 11 examples of data set *B*
    5. Construct “classifier model” Ma on data set A & Mb on data set B
    6. Record which one is better
    7. Add noise to obtain ‘estimated’ probabilities

       $$p' = p + k * U(-0.5, 0.5)$$
    8. Which one is better now

Experimental Result 1



- AUC, Brier score and Accuracy are more vulnerable to the existence of noise in the predicted probabilities
- the model selected by sAUC more reliable

Experiment 2

- *17 real data sets selected from the UCI repository*
  - 11 small data sets
    - Training data 50%
    - Validation data 10%
    - Test data 40%
  - 6 larger data sets
    - Training data 50%
    - Validation data 25%
    - Test data 25%

- Train 10 different classifiers with the same learning technique

  (J48, Naive Bayes, and Logistic Regression) over the same training data, by randomly removing three attributes before training
- Model selected according to AUC, Brier Score and sAUC

Experimental Result 2

#	J48 sAUC	J48 AUC	J48 BS	Naive Bayes sAUC	Naive Bayes AUC	Naive Bayes BS	Logistic Regression sAUC	Logistic Regression AUC	Logistic Regression BS
1	86.34	**83.76**	**85.81**	70.80	**67.98**	**69.96**	70.07	**67.28**	**69.23**
2	51.79	**51.32**	**51.05**	51.19	51.81	51.78	51.19	51.76	51.80
3	95.92	**93.20**	**95.47**	95.47	**92.21**	**94.96**	95.98	**92.65**	**95.58**
5	79.48	**77.72**	**78.16**	72.13	**70.88**	**71.05**	74.62	**72.11**	**72.68**
6	90.16	**89.25**	**89.56**	89.70	**89.06**	**89.61**	91.12	**90.62**	**90.55**
7	68.95	**68.75**	**68.85**	77.69	**77.24**	**77.25**	77.60	**77.29**	**77.20**
9	98.11	**97.81**	**97.98**	96.90	**96.74**	**96.81**	98.36	**98.24**	**98.28**
10	61.75	62.10	62.09	69.62	**69.09**	**68.98**	65.19	**64.94**	65.33
11	97.68	**97.64**	**97.67**	98.01	**97.94**	**98.00**	99.24	**99.18**	**99.22**
12	87.13	**85.65**	**86.13**	83.85	**83.60**	**83.82**	84.18	**83.74**	**83.76**
13	83.42	83.56	83.45	88.69	**88.68**	**88.49**	89.24	**89.12**	**89.13**
wins		9	9		10	10		10	9

#	J48 sAUC	J48 AUC	J48 BS	Naive Bayes sAUC	Naive Bayes AUC	Naive Bayes BS	Logistic Regression sAUC	Logistic Regression AUC	Logistic Regression BS
4	99.92	**99.91**	**99.91**	95.88	96.45	96.45	99.59	**99.55**	**99.57**
8	96.69	96.78	**96.67**	95.88	96.50	96.45	96.95	**96.93**	**96.91**
14	98.70	**98.67**	**98.65**	91.85	92.00	**91.62**	93.68	93.78	**93.59**
15	69.55	69.67	69.90	70.47	70.59	70.75	94.83	96.55	94.90
16	96.73	97.28	**96.59**	98.00	**97.99**	**97.90**	96.91	97.01	96.98
17	100	100	100	99.80	99.88	**99.79**	100	100	100
wins		2	3		1	3		2	3

- The performance of each selected classifier model is accessed by AUC on the test data

- sAUC is a good classifier model selector when the validation data is limited

Conclusions

- When the validation data is limited, only ROC curve & AUC is not enough to evaluate the performance of scoring classification models due to overfitting the validation data
- This paper mainly studied "*how quickly AUC deteriorates if the positive scores are decreased*"
- The concept of sROC curve & sAUC is presented, which uses both ranking information and score information
- The problem of overfitting can be avoided effectively

Thank you