Evaluation of learning outcomes

COMP 135 Intro to Machine Learning
Liping Liu
slides modified from Roni Khardon’s with permission

Evaluating machine learning outcomes

• More concrete questions:
• What quantity should we measure?
• How can we estimate it?
• Can we get quantitative guarantees for estimates and comparisons?

Evaluating machine learning outcomes

Evaluation measure - accuracy

• Suppose the classifier predict $\hat{y}_i$ while the true label is $y_i$
• Accuracy
  \[ accuracy = \frac{[y_i = \hat{y}_i]}{N} \]
• Problem:
  - How about the test set has 990 negatives and 10 positives while the classifier always predict negative?

Evaluation measure – error rate

• Suppose the classifier predict $\hat{y}_i$ while the true label is $y_i$
• Error rate
  \[ error rate = \frac{[y_i \neq \hat{y}_i]}{N} = 1 - accuracy \]
• Same problem as accuracy

Evaluation measure – average cost

• Suppose the classifier predict $\hat{y}_i$ while the true label is $y_i$
• Average cost
  \[ cost = \frac{TP \cdot C_{11} + FP \cdot C_{01} + FN \cdot C_{10} + TN \cdot C_{00}}{N} \]

Evaluation measure - ROC

• "Give me values before thresholding"
  - Sense of ranking in some applications
  - E.g. choose top K positives
• Plot measures against different threshold values
  - ROC curve
    \[ TPR = \frac{TP[y_i = 1 \text{ and } \hat{y}_i = 1]}{TP[y_i = 1]} \]
    \[ FPR = \frac{FP[y_i = 0 \text{ and } \hat{y}_i = 1]}{FP[y_i = 0]} \]
Supervised Learning

Application → New Data

Training Data → Learning Algorithm → Classifier

Predictions of Labels for new data

In real application

Training Data → Unknown Test data

Classifier

Training accuracy → Test accuracy

What is my real accuracy?

In real application

Training Data → Validation Data → Unknown New Data ??

Classifier

Training accuracy → Validation accuracy → Test accuracy

Good approximation of test accuracy

Evaluating a classifier

NEVER touch the validation data when training a classifier!!

Validation set

- Validation Set Method:
  - keep aside a portion of the example set
  - Train model on remaining data
  - Measure performance on validation set

- (+) Unbiased estimate of quantity
- (-) Wastes data ...

Cross validation

- Cross Validation Method:
  - Divide data into k portions (called folds)
  - Repeat for i in {1,2,...,k}:
    - Train model on data from all folds except i
    - Measure performance on i'th fold
  - Average the performance

- Test data in different folds is disjoint
- But training sets are not
Stratified Cross Validation

- Test folds do not have the same class frequencies - will increase variance
- Solution: stratify samples of each class across folds
- Cross Validation has become the standard and most widely used method for estimation of test performance

Leave One Out Method

- Setting k=N (number of examples) we get the leave one out method
- Effective in performance estimation
- Expensive to train classifiers N times
  - Not true for KNN classifier

Model Selection

- Model selection for test data

<table>
<thead>
<tr>
<th>Models</th>
<th>Validation accuracy</th>
<th>Your choice?</th>
</tr>
</thead>
<tbody>
<tr>
<td>KNN, K=1</td>
<td>0.85</td>
<td></td>
</tr>
<tr>
<td>KNN, K=3</td>
<td>0.89</td>
<td></td>
</tr>
<tr>
<td>KNN, K=5</td>
<td>0.87</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Models</th>
<th>Validation accuracy</th>
<th>Your choice?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Reg., λ = 0.1</td>
<td>0.92</td>
<td></td>
</tr>
<tr>
<td>Logistic Reg., λ = 1</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>Logistic Reg., λ = 10</td>
<td>0.90</td>
<td></td>
</tr>
</tbody>
</table>

Quantitative Comparisons

- Can you claim that LR with λ = 1 is better than KNN with K=3?
  
  NO

<table>
<thead>
<tr>
<th>Models</th>
<th>Validation accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>KNN, K=3</td>
<td>0.85 ± 0.1</td>
</tr>
<tr>
<td>Logistic Reg., λ = 1</td>
<td>0.95 ± 0.1</td>
</tr>
</tbody>
</table>

Quantify uncertainty of error rate

- Wrap a classifier as a black box
  - However fancy it is - not related to the variance of it’s error rate
- Errors as samples from a Bernoulli distribution
  - the true error rate is the probability

<table>
<thead>
<tr>
<th>Test inst</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>error/cost</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Variance

- Error rate estimation: \( \hat{\epsilon} = 0.25 \)
  - \( \hat{\epsilon} \) is approximately Gaussian-distributed when \( N \) is large. Why?
- Variance estimation
  \[
  \sigma = \text{std}(\epsilon) = \sqrt{\frac{N\epsilon(1-\epsilon)}{N-1}}
  \]

Hypothesis test

- \( H_0: \) the true error rate \( \epsilon \) is \( \epsilon_0 \)
  - Confidence interval \([\epsilon_0 - 2.365\sigma, \epsilon_0 + 2.365\sigma]\) with 95% confidence level
- \( H_0: \) the true error rate \( \epsilon \) is at least \( \epsilon_0 \)
  - Confidence interval \([\epsilon_0 - 1.895\sigma, 1]\) with 95% confidence level

Hypothesis test – model comparison

- \( H_0: \) method 1 does not perform better than method 2.

\[
\begin{array}{cccccccc}
\text{test/test} & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\
\text{error/cost 1} & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
\text{error/cost 2} & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
\text{error diff} & -1 & 1 & 0 & 0 & -1 & 0 & 0 & 0 \\
\end{array}
\]

- Equivalently, test whether the error difference is zero

Evaluating machine learning outcomes

- You should now have some answers to these questions:
  - What quantity should we measure?
  - How can we estimate it?
  - Can we get quantitative guarantees for estimates and comparisons?