How to evaluate machine learning

Problem based learning - 28. October 2019
How would you measure performance for a given predictor?
Performance measurements (1)

- **Binary classification (2 classes):**
  - True Positive: “Positive” sample, predicted as “positive”
  - False Positive: “Negative” sample, predicted as “positive”
  - True Negative: “Negative” sample, predicted as “negative”
  - False Negative: “Positive” sample, predicted as “negative”

- **Clearly state what you define as the positive class!**

- **General case (n classes): more complicated**
  - “Positives”: Sample belongs to class i and is predicted in class i, for all classes i
  - “Negatives”: Sample belongs to class i and is not predicted in class i, for all classes i
  - Make n binary comparisons

→ Confusion matrix to show your classification
### Performance measurements (2)

**Confusion matrix**

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class 1</td>
<td>39</td>
<td>6</td>
<td>5</td>
<td>50</td>
</tr>
<tr>
<td>Class 2</td>
<td>1</td>
<td>42</td>
<td>7</td>
<td>50</td>
</tr>
<tr>
<td>Class 3</td>
<td>7</td>
<td>13</td>
<td>30</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>47</td>
<td>61</td>
<td>42</td>
<td>150</td>
</tr>
</tbody>
</table>

**Normalized confusion matrix**

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class 1</td>
<td>0.78</td>
<td>0.12</td>
<td>0.1</td>
<td>50</td>
</tr>
<tr>
<td>Class 2</td>
<td>0.02</td>
<td>0.84</td>
<td>0.14</td>
<td>50</td>
</tr>
<tr>
<td>Class 3</td>
<td>0.14</td>
<td>0.26</td>
<td>0.6</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>47</td>
<td>61</td>
<td>42</td>
<td>150</td>
</tr>
</tbody>
</table>
Performance measurements (2)

TP: Correctly predicted as positive
FP: Falsely predicted as positive
TN: Correctly predicted as negative
FN: Falsely predicted as negative

Precision (Accuracy) = TP/(TP+FP)
Coverage (Recall) = TP/(TP+FN)
F1 = 2*Cov*Prec/(Cov+Prec)

Neg. Precision = TN/(TN+FN)
Neg. Coverage = TN/(TN+FP)

Q(n) (Overall accuracy) =
(TP+TN)/(TP+FP+TN+FN) =
#correctly predicted/#all samples

MCC=(TP*TN-FP*FN)/√[(TP+FP)(TP+FN)
(TN+FP)(TN+FN)]
Performance measurements (3)

Precision-Recall-Curve

ROC-Curve
Comparison to random/other methods

- To assess whether your model’s performance is significant:
  - compare performance to random baseline
  - compare performance to pre-existing method
  - provide standard errors: differences are only significant if they are not within the standard error (for now…)

- Random baseline, e.g.
  - take the class distribution for your dataset and assign labels at random as a prediction
  - assign everything to the majority class

- Never state more digits than supported by the standard error
  - e.g. Q18=0.796853±0.001 → Q18=0.797±0.001
How to calculate standard errors

● Per-residue prediction:
  ○ Calculate any performance measurement for each protein
  ○ → distribution of values
  ○ SE = SD/√(n-1) where n = number of proteins and SD=standard deviation

● Per-protein prediction:
  ○ you only have one value, no distribution
  ○ estimate the underlying distribution using bootstrapping:
    ■ randomly draw $n$ samples with replacement from your original samples
      (n= number of samples)
    ■ calculate performance measurement
    ■ repeat $m$ times (e.g. 1000 times) → $X= (x_1, \ldots, x_m)$
    ■ SE=SD($X$)
Any questions?
Tips on scientific paper writing

Problem based learning - 28. October 2019
Content

1. Requirements for final report
2. Content/Structure of the final report
3. General remarks on scientific writing
Requirements for final report

- use the Bioinformatics template (link on the course website)
- use either Word or Latex
- submit the final report as PDF
- write ~5 pages
Content/Structure of the final report

Major parts:

1. Abstract
2. Introduction
3. Material & Methods
4. Results & Discussion
5. Conclusion
Abstract

What should be part of the abstract?
Abstract

Abstract should cover:

● why is the work important?
  ○ what questions were you trying to answer?
  ○ what are the most significant implications?
● short overview of the main methodology
● summary of the major results (use numbers + standard errors)

Should be concise and readable, approx. 400 words

Do not include references
Introduction

What should be part of the introduction?
Introduction

Introduction should cover:

● (biological) background to understand
  ○ your work
  ○ the motivation behind your work

● existing methods/state-of-the-art performance
  ○ on what previous work/assumptions is your work based?
  ○ no technical details
  ○ provide proper references

● objectives of your work

Only focus on aspects directly relevant for your work!
Material & Methods

What should be part of material & methods?
Material & Methods

Material & Methods should cover:

● the dataset you used and how it was constructed
● any pre-existing methods you applied
● technical details of your own implementation
● assessment and evaluation of your work

Do not state results here!

This section should focus on technical details and should explain all concepts you applied.
What should be part of the results & discussion?
Results & Discussion

Results:

- statement of observations (statistics, tables, figures)
- include negative results

Discussion:

- interpretation of results
- what can we learn from the results?
- what is the significance of the results?

Results and discussion can also be separated
Conclusion

What should be part of the conclusion?
Conclusion

Conclusion should cover:

● strongest and most important statement(s) from your observations
● broader implications of your results
● maybe outlook outlining further steps of the project

Do not repeat word for word the abstract, introduction or discussion
General remarks on scientific writing (1)

● include references in the text
  ○ everything that you take from another publication has to be cited within the text (i.e. after the statement you want to cite)
  ○ for every citation, include a reference at the end in the “References” section
  ○ citation manager: EndNote (Word), BibTex (LaTex), …

● include figures/tables in the text to show your results
  ○ figures and tables should be referenced in the text
  ○ provide a meaningful caption summarising the major result of the figure/table
  ○ make sure everything is readable (use a large enough font and colors that are distinguishable)
General remarks on scientific writing (2)

● don’t use colloquial terms
  ○ use do not, cannot,... instead of don’t, can’t,...
● try to avoid passive voice
● try to avoid empty phrases
  ○ use connecting words like therefore, furthermore,... only if really necessary
  ○ don’t use “Figure 1 shows that xy leads to z”, just write “xy leads to z (Fig. 1)”
● try to avoid too long sentences
Any questions?