An Exploration of Automated Grading of Complex Assignments
An Exploration of Automated Grading of Complex Assignments
An Exploration of **Automated Grading** of Complex Assignments
An Exploration of Automated Grading of Complex Assignments

M. Brooks et al. “Divide and Correct: Using Clusters to Grade Short Answers at Scale”. In: ACM L@S. Atlanta, Georgia, USA, 2014, pp. 89–98. ISBN: 9781450326698

A. Nguyen et al. “Codewebs: Scalable Homework Search for Massive Open Online Programming Courses”. In: WWW. 2014, pp. 491–502

C. P. Rosé et al. “A Hybrid Text Classification Approach for Analysis of Student Essays”. In: BEA. ACL, 2003, pp. 68–75
(Why) Do we need these kinds of assignments?
You are a practitioner with an interest in equine medicine. During a routine visit to an area stable, your client asks you to perform a physical examination and to draw blood and collect urine from a near weaning Thoroughbred foal for future sale. The potential buyer wants a routine examination before purchasing the animal. The foal is high spirited and makes the client chase him around the paddock a few times before he can be halted. No abnormalities were found on physical examination.

**HEMATOLOGY**

<table>
<thead>
<tr>
<th>Test name</th>
<th>Test Result</th>
<th>Ref. Int.</th>
<th>Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>RBC</td>
<td>13.1</td>
<td>6.0-12.0</td>
<td>n*10^6/ul</td>
</tr>
<tr>
<td>HGB</td>
<td>19</td>
<td>10.0-18.0</td>
<td>g/dl</td>
</tr>
<tr>
<td>HCT</td>
<td>52</td>
<td>32.0-48.0</td>
<td>%</td>
</tr>
<tr>
<td>MCV</td>
<td>39.7</td>
<td>34.0-58.0</td>
<td>fl</td>
</tr>
<tr>
<td>MCH</td>
<td>14.5</td>
<td>13.0-19.0</td>
<td>pg</td>
</tr>
<tr>
<td>MCHC</td>
<td>36.5</td>
<td>31.0-37.0</td>
<td>g/dl</td>
</tr>
<tr>
<td>NRBC</td>
<td>0</td>
<td>n/100 wbc</td>
<td></td>
</tr>
<tr>
<td>ANISO</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BILI</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**URINALYSIS (VOIDED)**

<table>
<thead>
<tr>
<th>Test name</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>COLOR</td>
<td>straw</td>
</tr>
<tr>
<td>TRANSP</td>
<td>clear</td>
</tr>
<tr>
<td>S. G.</td>
<td>1.026</td>
</tr>
<tr>
<td>pH</td>
<td>7.5</td>
</tr>
</tbody>
</table>
The Applied Learning Platform (ALP)

www.whenknowingmatters.com
The Applied Learning Platform (ALP)

www.whenknowingmatters.com
The Applied Learning Platform (ALP)

Todd, Weak and Lethargic Dog

What Endocrine Disease is Most Likely in this Case?

Answer: Addison’s disease or deficiency of glucocorticoid and mineralocorticoid

Historical and Physical Exam Findings that Support Endocrine Disease Chosen

Physiology: Associated with Mineralocorticoid Deficiency

- hypovolemia, hyponatremia, hyperkalemia, dehydration, and shock
- recumbency and weakness
- anorexic and lethargic
- episodes of vomiting foamy bile
- weak on presentation and appeared very depressed
- MM were tacky
- CRT was >2 seconds
- femoral pulses were weak
- 8% dehydrated

Initially the blood pressure was undetectable

- after a 2L bolus of fluids (0.9% saline), the pressure increased to 110 mmHg Quality Evidence

Physiology: Associated with Glucocorticoid Deficiency

- Associated with low cortisol concentrations; notice that there is overlap with volume and blood pressure effects of mineralocorticoids
- intermittent lethargy and decreased appetite

episodes of vomiting foamy bile

- scant tarry stools

glucocorticoids cause leakiness of g.i. blood vessels and loss of blood and protein into gut
lost weight recently

glucocorticoids are necessary for optimal fat depots and lipogenesis
The Rubric—Likert; 1 (novice)–5 (expert)

Developing relevant refining (or clarifying) questions to answer based upon an honest assessment of current knowledge base

Questions: 2.82 ± 0.68

Approach to seeking answers to developed questions—literature search, etc.

Answers: 3.03 ± 0.77

Judgment of Quality of Information—awareness and application of standards of a discipline, bias detection including appropriate humility to detect one’s own potential bias, application of statistical concepts

Quality: 3.11 ± 0.98

Analysis of an argument

Analysis: 2.64 ± 0.77

Clarity and communication (written or oral) of thought: conciseness, grammar, spelling, elocution

Clarity: 3.38 ± 0.94

Application and understanding of appropriate disciplinary content

Application: 2.87 ± 0.57

(Mean and standard deviation, $n = 107$)
1. **Feasibility**: How effective are state of the art machine learning approaches for automated grading? Are they sufficiently effective to be immediately useful in practice?

2. **Formulation and Evaluation**: What is the right way to formulate the grading problem as a machine learning problem? What is the right way to measure effectiveness?

3. **Integration**: How should an automated grader be integrated with manual instructor/TA grading? What are the trade-offs?
Feasibility of Automatic Grade Prediction

\[ MAE = \frac{1}{n} \sum_{i=1}^{n} |r(f(x_i)) - r(y_i)| \]

<table>
<thead>
<tr>
<th>Feature</th>
<th>Baseline ( \pm ) Standard Error</th>
<th>SVOR(^1 ) ( \pm ) Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \text{sim} ) (3)</td>
<td>0.9358 ( \pm ) 0.0882</td>
<td>0.8811 ( \pm ) 0.0940</td>
</tr>
<tr>
<td>( \text{sim + sel} ) (5)</td>
<td>0.9566 ( \pm ) 0.1677</td>
<td>0.8642 ( \pm ) 0.0325</td>
</tr>
<tr>
<td>( \text{toks} ) (2646)</td>
<td>0.9075 ( \pm ) 0.0789</td>
<td>0.7660 ( \pm ) 0.0910(^\dagger)</td>
</tr>
<tr>
<td>( \text{all} ) (2651)</td>
<td>0.9792 ( \pm ) 0.1568</td>
<td>0.7566 ( \pm ) 0.0738(^\dagger)</td>
</tr>
</tbody>
</table>

\(^\dagger\): statistically significant using an unpaired \( t \)-test with \( p \leq 0.05 \).

**Table**: Effectiveness (in terms of MAE) of incorporating additional features in grade prediction for “quality” dimension using SVOR methods compared to the mode-assigning baseline. Number of features is given in parenthesis.

1[^1]: [http://www.work.caltech.edu/~htlin/program/libsvm/](http://www.work.caltech.edu/~htlin/program/libsvm/)
Feasibility of Automatic Grade Prediction

Table: Similar experiment to Table 1, but for “clarity” dimension.

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>SVOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>sim (3)</td>
<td>0.7906 ± 0.0771</td>
<td>0.7830 ± 0.0836</td>
</tr>
<tr>
<td>sim + sel (5)</td>
<td><strong>0.7623 ± 0.0649</strong></td>
<td>0.7811 ± 0.0561</td>
</tr>
<tr>
<td>toks (2646)</td>
<td>0.7528 ± 0.0550</td>
<td><strong>0.7415 ± 0.0597</strong></td>
</tr>
<tr>
<td>all (2651)</td>
<td><strong>0.7189 ± 0.0617</strong></td>
<td>0.7226 ± 0.0527</td>
</tr>
</tbody>
</table>

- **Main takeaway:** effectiveness very much depends on
  1. rubric dimension
  2. features used
Is outright grade prediction really our goal? 
(Hint: maybe it shouldn’t be)

1 Annotator agreement?
   - Low in practice\(^2\), *even for short-answer questions!*
   - Only going to be worse for complex assignments...
   - **Significant** barrier to leveraging peer grading

2 The machine will always be imperfect.
   - How sensitive is the grading process to “small” mistakes?
   - Can we reduce this sensitivity?

Can we treat grading as a **ranking problem?** (Hint: probably, or I wouldn’t put it on a slide)

1 Annotator agreement?
   - **Much easier** to get people to agree on “is \(a\) better than \(b\)?”
   - Provides an easy mechanism for **leveraging peers**: their inherent positivity (bless them) **won’t bias a pairwise decision**!

2 The machine will always be imperfect.
   - **Significantly less sensitive** to minute mistakes in the ranking, as long as it is largely consistent with instructor preference
   - Grading \(\rightarrow\) assigning cutoffs: quite natural!

**Evaluation metric**: “distance” between machine and instructor ranking (lower is better)

\[
NDPM = \frac{2n_d + t_x}{2(n_c + n_d + t_x)}
\]

---


Is **batch-mode machine learning** the right setup?  
(Hint: maybe it shouldn’t be)

- **Batch-mode learning**: Grade a bunch of assignments (or provide a bunch of pairwise assessments), give them to the machine, wait for a bit, and use its output on unlabeled assignments for grading
  - This is **not very collaborative**!
  - Machine is not able to inform instructor to grade *most helpful assignments/pairs*!
  - Unclear stopping decision: how many more should I grade to get a certain level of accuracy?
Can **active learning**\(^5\) be more effective?
(Hint: probably, or I wouldn’t put it on a slide)

**Active learning:**

1. Grade some small number of assignments/pairs
2. Provide them to the machine to learn from
3. **The machine suggests** the next assignment/pair to grade
4. (Go to 1 until you’re satisfied with the machine’s output)

 benefits:

- This is **collaborative by design**
- A clear stopping criterion: after each assignment/pair a new ranking is generated—stop when you’re happy with it

\(^5\) no, not that kind of active learning
Figure: A comparison between a randomized learning solution and an active learning solution to the grading-as-ranking problem. Reported is the average NDPM (lower is better) over 5 runs, with error bars indicating one standard deviation.
Questions?

(Thanks!)