

An Exploration of Automated Grading of Complex Assignments

2016-04-26

Chase Geigle, ChengXiang Zhai, Duncan Ferguson

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?

The Applied Learning Platform (ALP)

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 haltered. No abnormalities were found on physical examination.

HEMATOLOGY

| Test name | Test Result | Ref. Int. | Units |
|-----------|-------------|-----------|-----------------------|
| RBC | 13.1 | 6.0-12.0 | n*10 ⁶ /ul |
| HGB | 19 | 10.0-18.0 | g/dl |
| HCT | 52 | 32.0-48.0 | % |
| MCV | 39.7 | 34.0-58.0 | fl |
| MCH | 14.5 | 13.0-19.0 | pg |
| MCHC | 36.5 | 31.0-37.0 | g/dl |
| NRBC | | 0 | n/100 wbc |
| ANISO | | | |
| PLT COUNT | | | |



URINALYSIS (VOIDED)

| | | |
|--------|-------|---------|
| COLOR | straw | |
| TRANSP | clear | |
| S. G. | 1.026 | |
| pH | 7.5 | 6.5-9.0 |

The Applied Learning Platform (ALP)

Show help

Presenting information

Build your formulation

Entire Case - Copy to clipboard

Presenting Information for: Todd, ein schwacher lethargischer Hund

Relevant Observations

Hochschule für sinnvoll ersicht.

KLINISCHE ALLGEMEINUNTERSUCHUNG:

Allgemein: ruhig, schwach, apathisch

Gewicht: 38,5 kg Body Condition Score: 4/5

Herzschlagfrequenz: 140 T=100.3°F Atemfrequenz: 20 Atemzüge/Minute.

Schleimhaut: dunkel pigmentiert; die Farbe war schwer zu bewerten. Trocknen und kapillare Füllungszeit war >2 Sekunden.

Ohren/Augen/Nase/Hals: unauffällig Die Untersuchung der Maulhöhle ergab, dass die Zähne mgr. Zahnbelag aufweisen und Halitose.

Lymphknoten: unauffällig

Abdomen: unauffällig und nicht dolent

Herz/Lungen: Thorax auskultiert keine Veränderungen. Aber die Herzfrequenz war erhöht und die Pulsqualität war schwach. Es gibt kein Herzgeräusch und keine Arrhythmie.

Rektales Examen: Das Colon war leer, nicht dolent; keine Menge palpirt; Der Kot war mäßig, die Kotfarbe war grün-braun

Neurologische Untersuchung: Unauffällig, ausser, dass Toby schwach war und nicht gern aufstehen konnte.

Haut: Unauffällig, aber circa 8% Dehydration.

Schmerz: Von 0-10 (höchste), Toby war 1/10.

| | | | |
|-----------------------------|------|-------------------|------|
| Lactat | 1 | <2.5 (mmol/L) | Obs> |
| Harnstoff | 34 | 6-30 (mg/dl) | Obs> |
| Kreatinin | 1.3 | 0.5-1.5 (mg/dl) | Obs> |
| Anionenlücke | 8.1 | 8-25 (mmol/L) | Obs> |
| Osmolalität | 277 | 280-310 (meq/L) | Obs> |
| pH | 7.40 | 7.35-7.45 (units) | Obs> |
| HCO3- | 15 | 16-24 (mmol/L) | Obs> |
| Hematokrit | 61 | 35-52(%) | Obs> |
| Hämoglobin | 20.3 | 12-18 (g/dl) | Obs> |
| Vor-ACTH cortisol | 20.3 | 58-144(nmol/L) | Obs> |
| Nach-ACTH cortisol (60 min) | 20.3 | 225-425 (nmol/L) | Obs> |

☐ Vermehrtes Liegen und Schwäche

☐ schlapp und lustlos

☐ Seit ein paar Monaten ist er krank

☐ intermittierende Lethargie und verminderten Appetit

☐ erbricht seit 2 Tagen Galle

☐ beerfarbenden Kot

☐ Gewicht verloren

☐ ausgeschlossen, dass Todd etwas giftig gefressen hat.

☐ [Na+] war 123 meq/liter and [K+] war 6.1 meq/liter.

☐ Tierarzt hat 500 ml 0.9% NaCl subkutan gegeben

☐ ruhig, schwach, apathisch

☐ kapillare Füllungszeit war >2 Sekunden

☐ 8% Dehydration

☐ Blutdruck: Zuerst war dieser nicht messbar

☐ nach 2L Flüssigkeit (0.9% NaCl-Lösung), stieg der Blutdruck bis 110 mmHg an

☐ - Na+ 135 141-152 (mmol/L)

☐ - K+ 5.6 3.9-5.5 (mmol/L)

☐ - Glucose 87 68-126 (mg/dl)

☐ - HCO3- 15 16-24 (mmol/L)

☐ - Vor-ACTH cortisol 20.3 58-144(nmol/L)

☐ - Nach-ACTH cortisol (60 min) 20.3 225-425 (nmol/L)

Sponsor(s), Author(s), Contributor(s), and Copyright

Prof Duncan C. Ferguson, VMD, PhD
JProf. Dr. med. vet. Marion Piechotta, JProf. Dr. med. vet. Marion Bankstaht, and Eric (Rick) M. Mills DVM, PhD, MSW

www.whenknowingmatters.com

The Applied Learning Platform (ALP)

Show help **Presenting information** **Build your formulation** **Reference formulation** **Entire Case - Copy to clipboard**

List of Abbreviations

Todd, Weak and Lethargic Dog

- ☐ SIGNALMENT: "Todd", 4.5 yo MC Labrador Retriever, 38.5 kg
- ☐ recumbency and weakness
- ☐ sternal recumbency

What Endocrine Disease is Most Likely in this Case?

Answer

Hypoadrenocorticism

Historical and Physical Exam Findings that Support Endocrine Disease Chosen

Physiology

- ☐ anorexic and lethargic
- ☐ intermittent lethargy and decreased appetite
- ☐ scant tarry stools over the past couple of days and has lost weight recently
- ☐ weak on presentation and appeared very depressed
- ☐ d appeared very depressed.
- ☐ HR: 140 T=100.3°F Resp Rate: 20 breaths/minute
- ☐ MM were tacky and CRT was >2 seconds
- ☐ heart rate was increased and femoral pulses were weak
- ☐ stool was greenish brown
- ☐ weak to rise and ambulate
- ☐ vomited 2 days ago, and has had several more episodes of vomiting foamy bile since then

Historical and Physical Exam Findings that Do Not Support Endocrine Disease Chosen

Tests that Support Endocrine Disease Chosen

- ☐ administered 500 ml 0.9% NaCl SQ
- ☐ bloodwork was performed

SIGNALMENT: "Todd", 4.5 yo MC Labrador Retriever, 38.5 kg

- ☒ recumbency and weakness
- ☒ anorexic and lethargic
- ☒ intermittent lethargy and decreased appetite
- ☒ scant tarry stools over the past couple of days and has lost weight recently
- ☒ bloodwork was performed
- ☒ abnormality noted was plasma [Na+] was 1...
- ☒ administered 500 ml 0.9% NaCl SQ
- ☒ weak on presentation and appeared very d...
- ☒ d appeared very depressed.
- ☒ HR: 140 T=100.3°F Resp Rate: 20 breaths...
- ☒ MM were tacky and CRT was >2 seconds
- ☒ heart rate was increased and femoral pulse...
- ☒ stool was greenish brown
- ☒ weak to rise and ambulate

Grouped Problems for Therapy

- ☐ Drug (generic name)
- ☐ Approach
- ☐ Class of Drugs
- ☐ Mechanism of Action
- ☐ Absorption, Administration, Adverse Effects
- ☐ Note Problems (Monitor)

Show presenting information **Formulation Only - Copy to clipboard** **Expand all entries** **Expand all notes**

www.whenknowingmatters.com

The Applied Learning Platform (ALP)

- F Todd, Weak and Lethargic Dog

- Q What Endocrine Disease Is Most Likely in this Case?

- A Answer: Addison's disease or deficiency of glucocorticoid and mineralocorticoid

- F Historical and Physical Exam Findings that Support Endocrine Disease Chosen

- P | Not e Physiology: Associated with Mineralocorticoid Deficiency

Note: hypovolemia, hyponatremia, hyperkalemia, dehydration, and shock

- O recumbency and weakness
- O anorexic and lethargic
- O episodes of vomiting foamy bile
- O weak on presentation and appeared very depressed
- O MM were tacky
- O CRT was >2 seconds
- O femoral pulses were weak
- O 8% dehydrated
- O Initially the blood pressure was undetectable

- O | Not e after a 2L bolus of fluids (0.9% saline), the pressure increased to 110 mmHg Quality Evidence

Note: consistent with low Na+ being very important to clinical presentation

- P | Not e Physiology: Associated with Glucocorticoid Deficiency

Note: Associated with low cortisol concentrations; notice that there is overlap with volume and blood pressure effects of mineralocorticoids

- O | Not e intermittent lethargy and decreased appetite

Note: glucocorticoids help sustain blood volume physiologically and also stimulate appetite (particularly in pharmacological quantities)

- O episodes of vomiting foamy bile

- O | Not e scant tarry stools

Note: glucocorticoids cause leakiness of g.i. blood vessels and loss of blood and protein into gut

- O | Not e lost weight recently

Note: 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, $n = 107$)

Research Questions

- 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

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

| | Baseline | SVOR ¹ |
|----------------------|-----------------|------------------------------------|
| sim (3) | 0.9358 ± 0.0882 | 0.8811 ± 0.0940 |
| sim + sel (5) | 0.9566 ± 0.1677 | 0.8642 ± 0.0325 |
| toks (2646) | 0.9075 ± 0.0789 | 0.7660 ± 0.0910[†] |
| all (2651) | 0.9792 ± 0.1568 | 0.7566 ± 0.0738[†] |

†: 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<http://www.work.caltech.edu/~htlin/program/libsvm/>

Feasibility of Automatic Grade Prediction

| | Baseline | SVOR |
|---------------|---------------------------------------|---------------------------------------|
| sim (3) | 0.7906 \pm 0.0771 | 0.7830 \pm 0.0836 |
| sim + sel (5) | 0.7623 \pm 0.0649 | 0.7811 \pm 0.0561 |
| toks (2646) | 0.7528 \pm 0.0550 | 0.7415 \pm 0.0597 |
| all (2651) | 0.7189 \pm 0.0617 | 0.7226 \pm 0.0527 |

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

□ **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 practice², *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?

²M. Mohler and R. Mihalcea. “Text-to-text Semantic Similarity for Automatic Short Answer Grading”. In: *EACL*. 2009, pp. 567–575.

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 ?”³
- 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 → assigning cutoffs: quite natural!

Evaluation metric: “distance” between machine and instructor ranking (lower is better)⁴

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

³C. Callison-Burch et al. “(Meta-) Evaluation of Machine Translation”. In: *WMT*. 2007, pp. 136–158.

⁴Y. Y. Yao. “Measuring Retrieval Effectiveness Based on User Preference of Documents”. In: *J. Am. Soc. Inf. Sci.* 46.2 (1995), pp. 133–145.

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

⁵no, not that kind of active learning

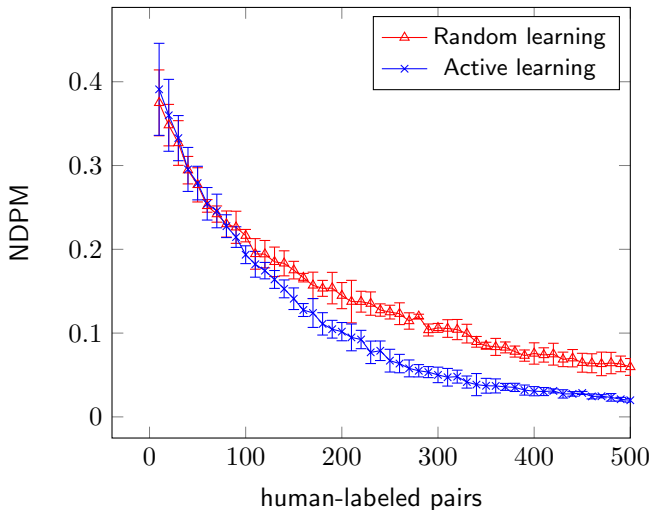


Figure: A comparison between a randomized learning solution and an active learning solution to the grading-as-ranking problem. Reported is the average NDPM (lower is better) over 5 runs, with error bars indicating one standard deviation.

Questions?

(Thanks!)