COMS30035, Machine learning: Combining Models 5, Ensembles of Humans

Edwin Simpson
edwin.simpson@bristol.ac.uk

Department of Computer Science, SCEEM
University of Bristol

November 16, 2023

Agenda

- Model Selection
- Model Averaging
- Ensembles: Bagging
- Ensembles: Boosting and Stacking
- Tree-based Models
- Conditional Mixture Models
- Ensembles of Humans

Wisdom of the crowd

- Remember the equation relating the error of a combination to the error of an average individual: $E_{COM} = \frac{1}{M} E_{AV}$
- Assuming uncorrelated, zero-mean errors
- Can we apply this when the base models are people rather than machine learners?
- Could combinations of humans be used in machine learning?

Dataset Annotation

- Annotation of datasets is extremely important for machine learning and scientific data analysis:
  - E.g., training labels for supervised learning
  - E.g., test labels for evaluation
  - Where do these annotations come from?

Dataset Annotation

- Annotation of datasets is extremely important for machine learning and scientific data analysis:
  - E.g., training labels for supervised learning
  - E.g., test labels for evaluation
  - Where do these annotations come from?
  - In a huge range of applications, somebody has to label the data manually (text, images, biological data, astronomy,...).


Dataset Annotation

- Annotation of datasets is extremely important for machine learning and scientific data analysis:
  - E.g., training labels for supervised learning
  - E.g., test labels for evaluation
  - Where do these annotations come from?
  - In a huge range of applications, somebody has to label the data manually (text, images, biological data, astronomy,...).

- There are many tasks that people can do that computers cannot, even though we make many errors.
  - E.g., following instructions
  - E.g., applying commonsense reasoning

Expert Annotators

- Expert annotators have a low average error, E_{AV}
- People make mistakes, even experts, so combine multiple annotations from different people.
- We would like to collect large datasets to support more extensive testing and more complex models
- Experts' time is expensive and limited, so how can we obtain large datasets at reasonable speed and cost?

Expert Annotators

- Expert annotators have a low average error, E_{AV}
- People make mistakes, even experts, so combine multiple annotations from different people.
- We would like to collect large datasets to support more extensive testing and more complex models
- Experts' time is expensive and limited, so how can we obtain large datasets at reasonable speed and cost?
- This is where the wisdom of the crowd comes in!

Crowdsourcing

- Ask a large number of non-expert annotators to provide the data!
- Crowdsourcing platforms allow *requesters* to create tasks for *crowd workers*

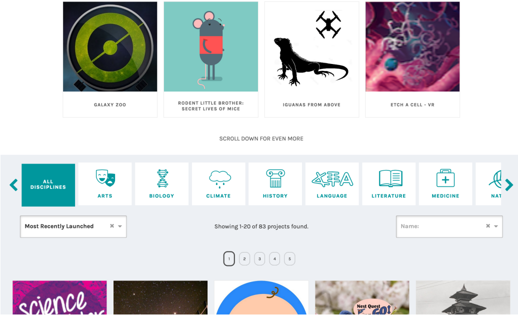
Crowdsourcing

- Ask a large number of non-expert annotators to provide the data!
- Crowdsourcing platforms allow *requesters* to create tasks for *crowd workers*
- Amazon Mechanical Turk – pay a few cents per task

Edwin Simpson
edwin.simpson@bristol.ac.uk

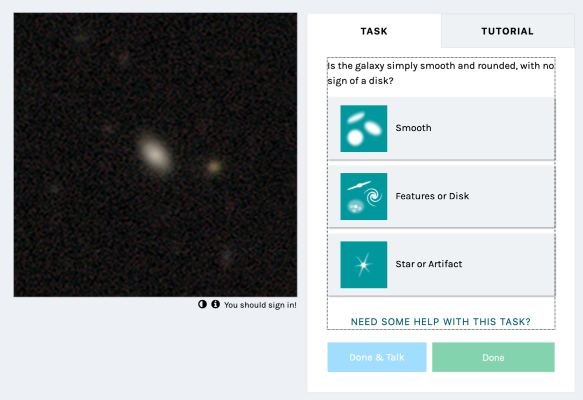
Crowdsourcing

- Ask a large number of non-expert annotators to provide the data!
- Crowdsourcing platforms allow *requesters* to create tasks for *crowd workers*
- Amazon Mechanical Turk – pay a few cents per task
- Zooniverse – volunteer citizen scientists



A Crowdsourcing Task

- Tasks need to be simple with clear instructions
- See http://www.zooniverse.org for many more



Crowd Size vs. Error Rate

- Crowdsourced annotations have lower quality, higher E_{AV}
- What can we do about this?

Crowd Size vs. Error Rate

- Crowdsourced annotations have lower quality, higher E_{AV}
- What can we do about this?
- Remember $E_{COM} = \frac{1}{M} E_{AV}$:
- We can prevent E_{COM} from rising by increasing M.
- So, using a larger number of crowd annotators allows us to obtain quality annotations at far reduced costs.

Example: Cheap and Fast – But Is It Good?

- Systematic comparison of workers to experts on various NLP tasks
- E.g. rate headlines to reflect emotional content

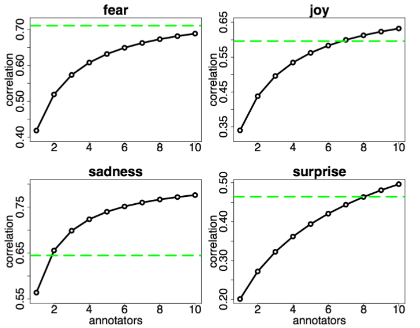
Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast - but is it good?: evaluating non-expert annotations for natural language tasks (EMNLP-08).

Example: Cheap and Fast – But Is It Good?

- Systematic comparison of workers to experts on various NLP tasks
- E.g. rate headlines to reflect emotional content
- Y-axis: correlation with mean of experts
- Green dashed lines: 1 expert
- Black solid lines: increasing number of workers per task

Outcry at N Korea 'nuclear test'

(Anger, 30), (Disgust,30), (Fear,30), (Joy,0), (Sadness,20), (Surprise,40), (Valence,-50).



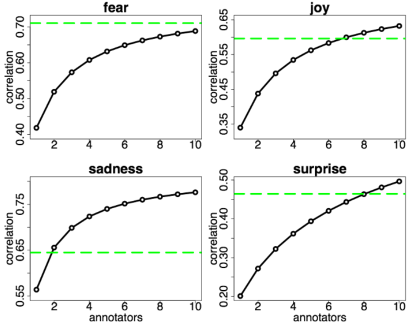
Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast - but is it good?: evaluating non-expert annotations for natural language tasks (EMNLP-08).

Example: Cheap and Fast – But Is It Good?

- Systematic comparison of workers to experts on various NLP tasks
- E.g. rate headlines to reflect emotional content
- Y-axis: correlation with mean of experts
- Green dashed lines: 1 expert
- Black solid lines: increasing number of workers per task
- On average, 4 workers compares to 1 expert

Outcry at N Korea 'nuclear test'

(*Anger, 30*), (*Disgust,30*), (*Fear,30*), (*Joy,0*), (*Sadness,20*), (*Surprise,40*), (*Valence,-50*).



Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast - but is it good?: evaluating non-expert annotations for natural language tasks (EMNLP-08).

Caveats

- Errors are not zero-mean and uncorrelated in practice.
- The design of the task, the way the data is presented to the crowd, the instructions and lack of expertise may mean that most annotators make the same mistakes.
- There are also spammers who don't make an effort to provide correct labels
- Annotators have different levels of skill

Remedies

- Correlated errors are tricky to deal with
- Address spamming and skill levels by learning a weighted combination function
- Similar to the *stacking* approach for ensembles of machine learners
- But not all annotators label all data points...
- Use the generative model for combining classifiers proposed by Dawid and Skene in 1979

Dawid, A. P., & Skene, A. M. (1979). Maximum likelihood estimation of observer error-rates using the EM algorithm. Journal of the Royal Statistical Society: Series C (Applied Statistics), 28(1), 20-28.

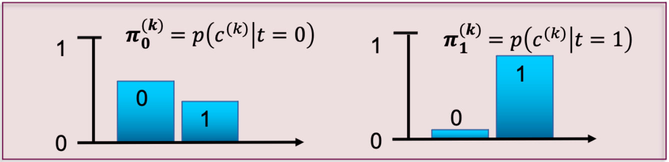
Dawid and Skene (1979)

- For each data point, each annotator k produces label $c^{(k)}$ from $\{1, ..., J\}$.
- To predict the true label t given a set of noisy annotations $\boldsymbol{c}$:

$$p(t = j|\boldsymbol{c}) = \frac{p(t = j) \prod_{k=1}^{K} p(c^{(k)}|t = j)}{\sum_{l=1}^{J} \left\{ p(t = l) \prod_{k=1}^{K} p(c^{(k)}|t = l) \right\}}. \tag{1}$$

Dawid and Skene (1979)

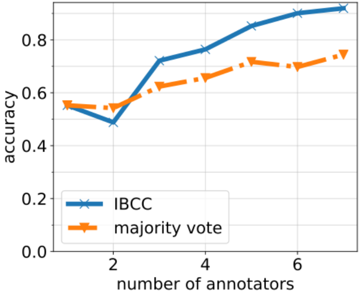
- Each annotator is modelled by a confusion matrix, $\boldsymbol{\pi}^{(k)}$ where each entry is $\pi_{ji}^{(k)} = p(c^{(k)} = i | t = j)$.
- $\boldsymbol{\pi}^{(k)}$ captures the annotator's different error rates for each class label.
- Can learn $\pi^{(k)}$ with EM, but Bayesian inference is more effective since the amount of data for many annotators is very small.



Edwin Simpson
edwin.simpson@bristol.ac.uk

Example of Dawid and Skene (simulated data)

Combining noisy classifications using majority vote vs Bayesian treatment of Dawid & Skene's model (IBCC).



Summary

- We can also combine people as well as machines using the same model combination principles.
- Combining lots of human annotators helps do tasks that machine learners can't do, such as constructing training and evaluation datasets.
- The Dawid and Skene model helps account for different skill levels and down-weight spammers and can be applied as a stacking approach to ensembles of machine learners.

Now do the quiz!

Please do the quiz for this lecture on Blackboard.