Machine Learning 2.04: Active Learning

Tom S. F. Haines
T.S.F.Haines@bath.ac.uk

UNIVERSITY OF BATH

Traditionally. . .

- Collect data set.
- Train model.
- Test model.
- Start using model.

Data sets

- Collecting data: Cheap
- Labelling: Expensive

- e.g. Doctors diagnosing x-rays (especially if segmenting).
- Can also be dangerous, e.g. invasive diagnostic procedures.

Unlabelled data

- Semi-supervised learning:
  Not all data is labelled.

Unlabelled data

- Semi-supervised learning:
  Not all data is labelled.

- Active-learning:
  Not all data is labelled...
  ...and the computer gets to decide what to label!

UNIVERSITY OF BATH

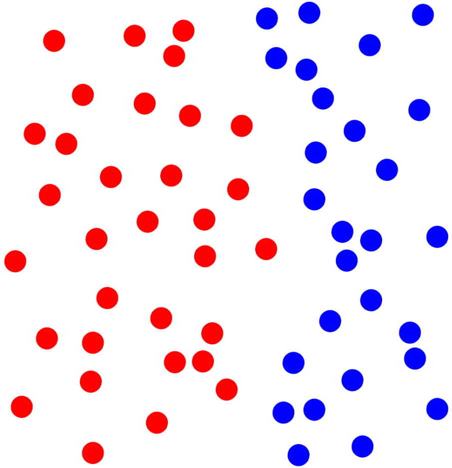
Unlabelled data

- Semi-supervised learning:
  Not all data is labelled.

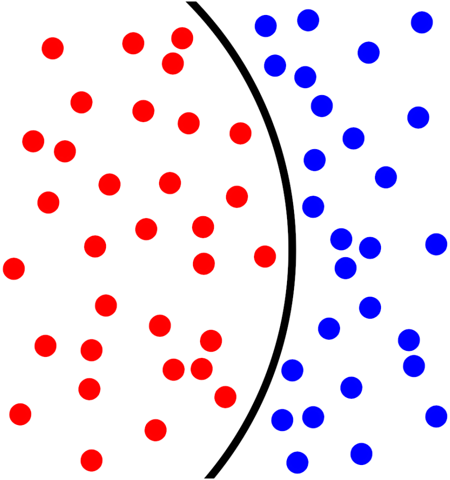
- Active-learning:
  Not all data is labelled...
  ...and the computer gets to decide what to label!

- Related to:
  - Optimal experimental design (statistics).
  - Automated science.
  - Hyperparameter optimisation.
  - Machine teaching.

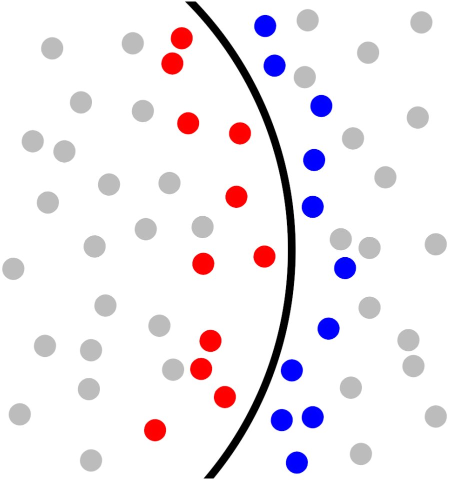
Labels are not equal



Labels are not equal

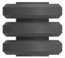


Labels are not equal

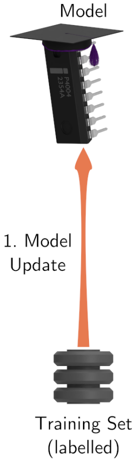


- Only need exemplars on edge! (same core idea as a support vector machine)

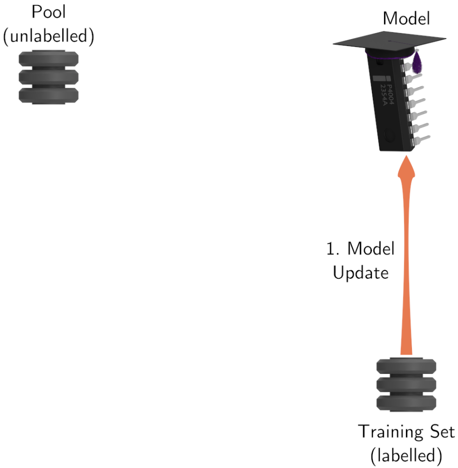
Active learning



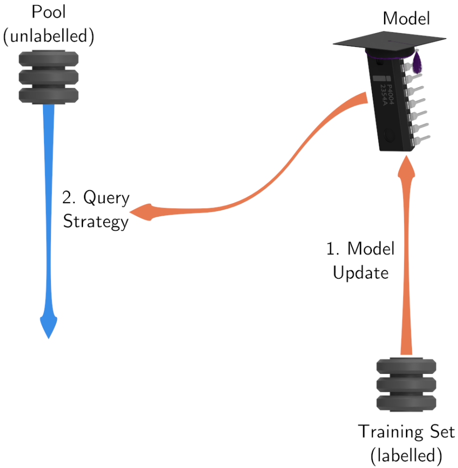
Active learning



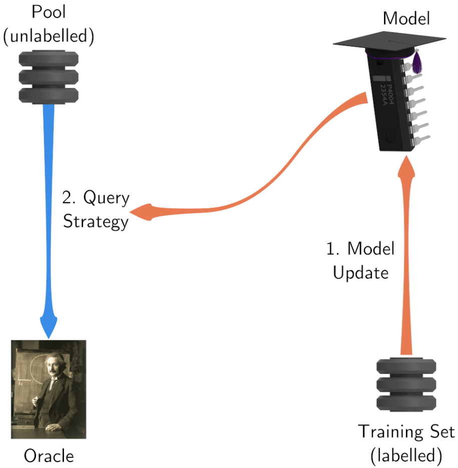
Active learning



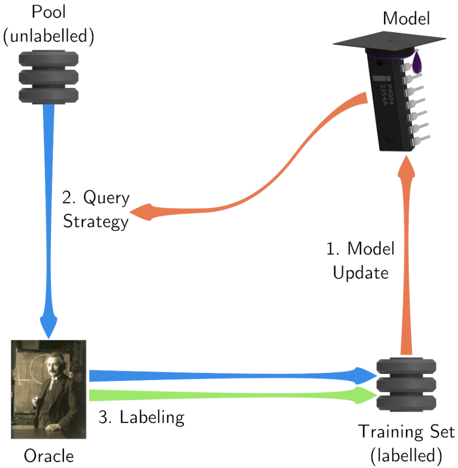
Active learning



Active learning



Active learning

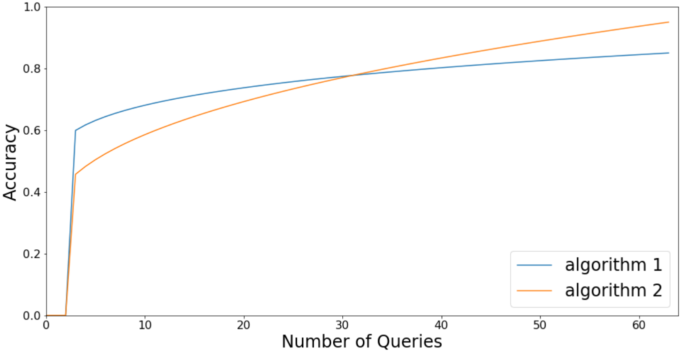


Measuring performance

- Goal: Maximum performance from least queries.

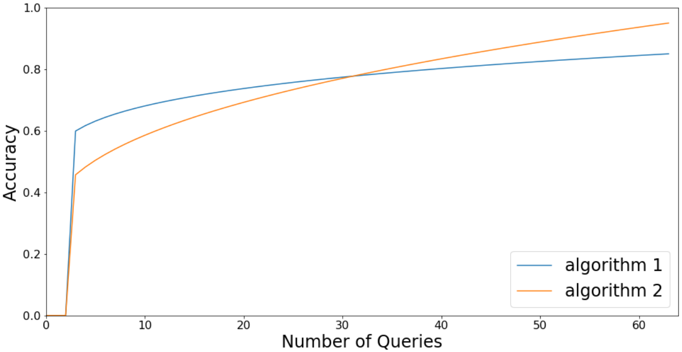
Measuring performance

- Goal: Maximum performance from least queries.



Measuring performance

- Goal: Maximum performance from least queries.



- Notes:
  - Fake the oracle when testing!
  - Average many runs for stochastic approaches.
  - Note that algorithms do crossover – stopping point may matter!
  - Want a number: Area under curve, or at specific point.

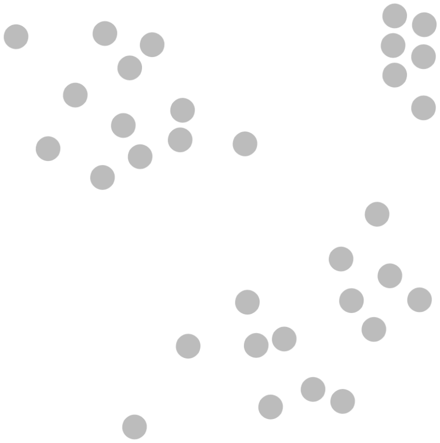
Stopping

- When to stop is usually problem specific.
- Running out of time or money.
- Improvement no longer seems worth it.
- Cost/benefit analysis.

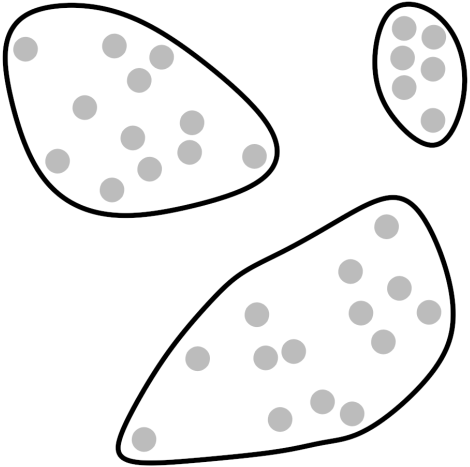
Baseline approach

- Randomly selecting an exemplar each time.
- Does much better than you would imagine! (on balanced data sets)

Clustering approach

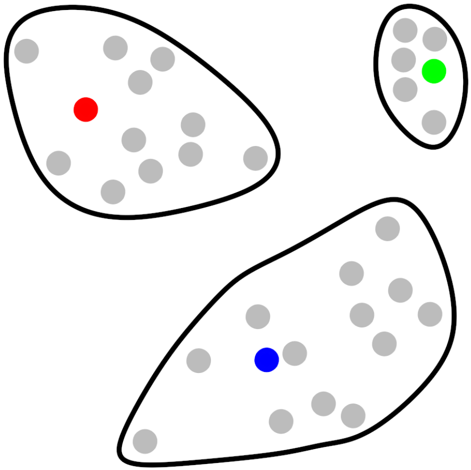


Clustering approach



- Cluster data.

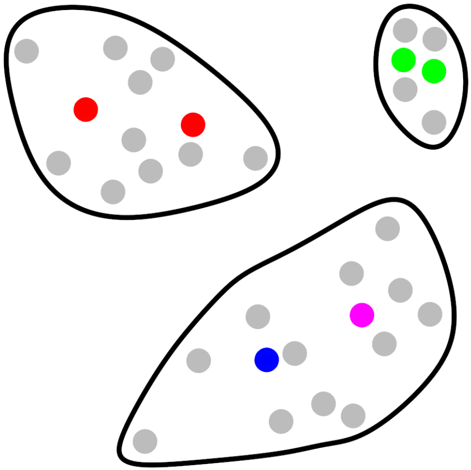
Clustering approach



- Cluster data.
- Query oracle for random exemplar in each cluster.

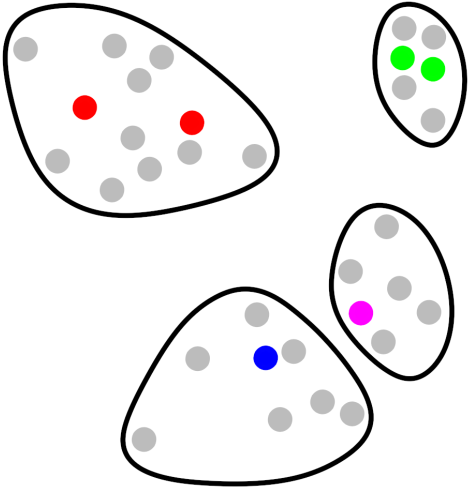
UNIVERSITY OF BATH

Clustering approach



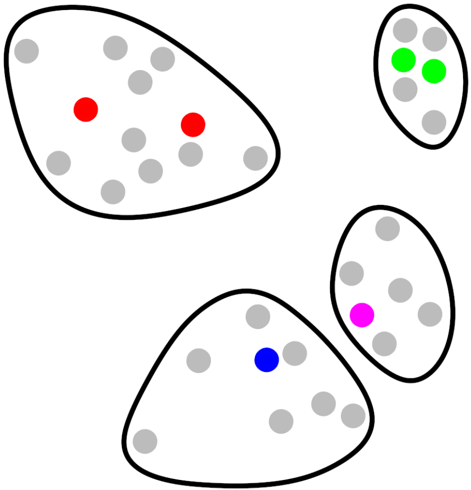
- Cluster data.
- Query oracle for random exemplar in each cluster.
- Repeat for confidence.

Clustering approach



- Cluster data.
- Query oracle for random exemplar in each cluster.
- Repeat for confidence.
- Contradiction $\implies$ refine clustering.

Clustering approach



- Cluster data.
- Query oracle for random exemplar in each cluster.
- Repeat for confidence.
- Contradiction $\implies$ refine clustering.

- Keep going until confident.
- Assumes classes are *separable*.

Probabilistic querying strategies

- Assume a probabilistic classifier.
  Can calculate $P(c|x)$ where $c \in C$ is a class for each exemplar, x.

- Approaches:

Probabilistic querying strategies

- Assume a probabilistic classifier.
  Can calculate $P(c|x)$ where $c \in C$ is a class for each exemplar, x.

- Approaches:
  - Uncertainty (margin) sampling: Exemplar with minimum $P(c_1|x) - P(c_2|x)$ where c_1 is the most probable class, c_2 the second most probable class.

Probabilistic querying strategies

- Assume a probabilistic classifier.
  Can calculate $P(c|x)$ where $c \in C$ is a class for each exemplar, x.

- Approaches:
  - Uncertainty (margin) sampling: Exemplar with minimum $P(c_1|x) - P(c_2|x)$ where c_1 is the most probable class, c_2 the second most probable class.
  - Least confident: Exemplar with minimum $\max_{c \in C} P(c|x)$.

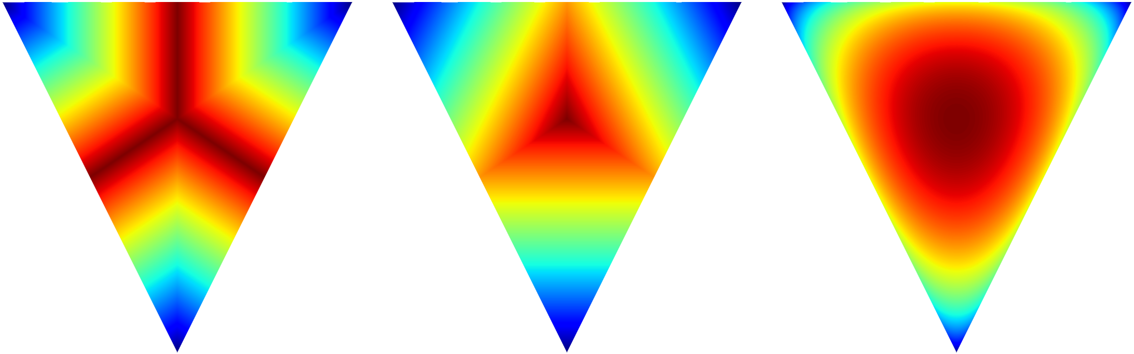
Probabilistic querying strategies

- Assume a probabilistic classifier.
  Can calculate $P(c|x)$ where $c \in C$ is a class for each exemplar, x.

- Approaches:
  - Uncertainty (margin) sampling: Exemplar with minimum $P(c_1|x) - P(c_2|x)$ where c_1 is the most probable class, c_2 the second most probable class.
  - Least confident: Exemplar with minimum $\max_{c \in C} P(c|x)$.
  - Entropy: Exemplar with maximum $-\sum_{c \in C} P(c|x) \log(P(c|x))$

Probabilistic querying visualisation

- Consider the intersection of three classes.
- Certain in the corners, blending over the triangle.



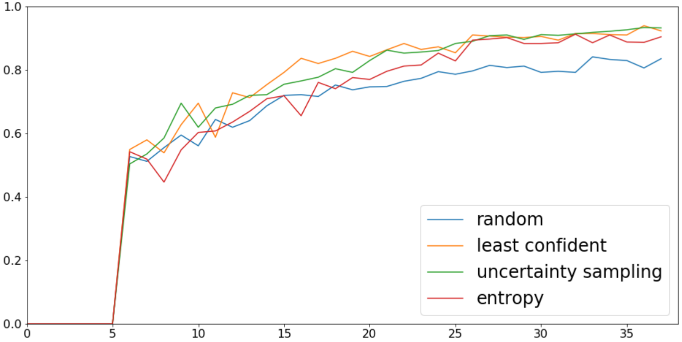
Uncertainty (margin) sampling Least confident Entropy

Example

- Problem: Identifying glass at a crime scene.
  Expensive because you have to break sheets of glass!

Example

- Problem: Identifying glass at a crime scene.
  Expensive because you have to break sheets of glass!



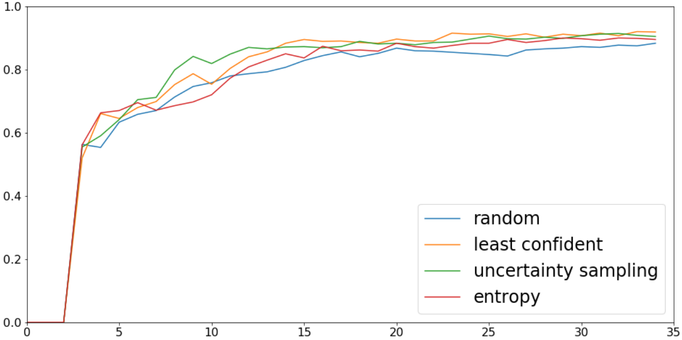
Random is competitive!

- Problem: Identifying (near-identical) seeds.
  Expensive because you have to wait for the plant to grow!

UNIVERSITY OF BATH

Random is competitive!

- Problem: Identifying (near-identical) seeds.
  Expensive because you have to wait for the plant to grow!



- Only for balanced data – real data is usually imbalanced.

Bias

- Active learning by design selects a biased sample.

Bias

- Active learning by design selects a biased sample.

- Classify entire pool to calculate class probabilities.
- Calibrate probability response (Kolmogorov–Smirnov test).

Query by committee

- Requires a Bayesian model.
- Draw multiple models from the posterior.
- Select exemplar with greatest disagreement.

- Slow.

Expected model change

- For each exemplar:
  1. Pretend it has been labelled by each class in turn.
  2. Measure how much the model parameters would change.
  3. Calculate expectation using current class assignment probabilities.

- Select exemplar with highest expected change.

- How to scale different parameters is unclear.
- Really slow.

UNIVERSITY OF BATH

Expected error reduction

- Simulate the results of multiple queries,
  e.g. update model with the next three queries, guessing the oracles answers.
- Estimate how much error is reduced with this extra information.
- Do every combination.
- Calculate expectation using current model probabilities.
- Select exemplar with best expected error reduction.

- Horrifically slow.

Time

- The previous approaches are slow.
- A human may be waiting.

Time

- The previous approaches are slow.
- A human may be waiting.

- Theoretically nice, pragmatically useless.

- Can optimise.
- Incremental learning is a given.

Active discovery

- Thus far: Assumed all classes known in advance.
- Often not true, e.g. only 0.001% of stars may be evidence of new physics.

Active discovery

- Thus far: Assumed all classes known in advance.
- Often not true, e.g. only 0.001% of stars may be evidence of new physics.

- Active discovery tries to find unknown classes.
- Typically outlier detection:
  Fit density estimate, give low probability exemplars to oracle.

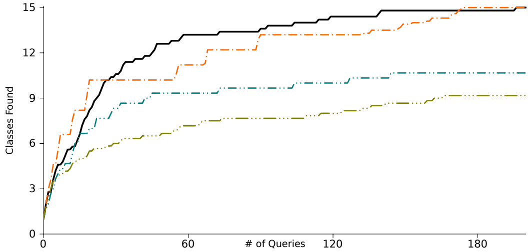
Active discovery and learning

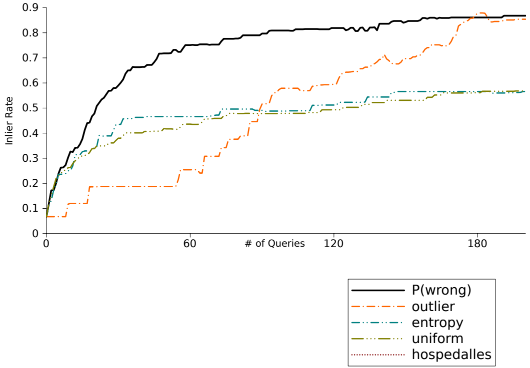
- Simultaneously:
    - Discover new classes.
    - Refine classification of existing.

- Exploration/exploitation trade off.

Dirichlet process approach

- Dirichlet process prior over the classes.
- Includes a probability for something new.
- Intrusion detection problem – KDD99.





Further variants

- Regression.

Further variants

- Regression.
- Batch: Select many exemplars at once.

UNIVERSITY OF BATH

Further variants

- Regression.
- Batch: Select many exemplars at once.
- Cost sensitive: Labelling cost varies; multiple labelling options
  (e.g. does the image contain a cat vs click on the cat vs segment the cat)

Further variants

- Regression.
- Batch: Select many exemplars at once.
- Cost sensitive: Labelling cost varies; multiple labelling options
  (e.g. does the image contain a cat vs click on the cat vs segment the cat)
- Online: Exemplars are seen once, and a decision to label has to be made immediately.

Further variants

- Regression.
- Batch: Select many exemplars at once.
- Cost sensitive: Labelling cost varies; multiple labelling options (e.g. does the image contain a cat vs click on the cat vs segment the cat)
- Online: Exemplars are seen once, and a decision to label has to be made immediately.
- Untrusted oracles: Oracles make mistakes (Amazon mechanical Turk).

Which approach?

- You can't run a proper experiment!
  - Identify and test similar problems.
  - Intuition for the rest!
  - Verify after the fact.
- Factor in needs of humans.

Summary

- Active learning: Computer asks for help.
- Active discovery: Computer finds new things.

- Many approaches.
- Be cautious!

Further reading

- Summary of active learning:
  "*Active Learning Literature Survey*",
  by B. Settles (2008).

Sources

- Glass identification:
  `https://archive.ics.uci.edu/ml/datasets/Glass+Identification`
- Seed identification:
  `http://archive.ics.uci.edu/ml/datasets/seeds`
- Intrusion detection:
  `http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html`