Lecture 16: Advanced Topics in Classification

William Webber (william@williamwebber.com)

COMP90042, 2014, Semester 1, Lecture 16

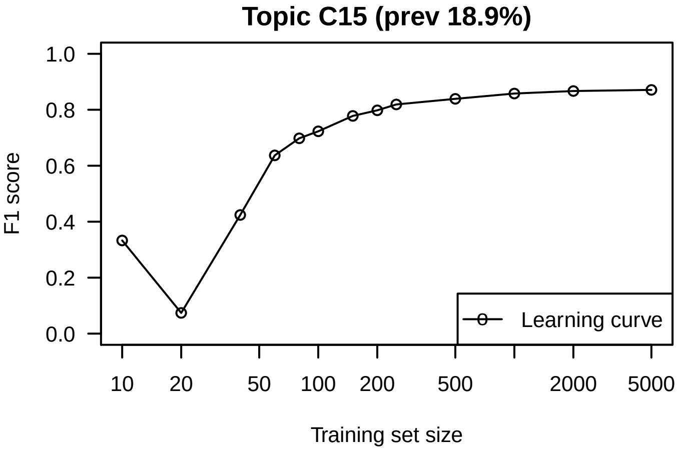
What we'll learn in this lecture

- Iterative training of classifier
- Calculation of learning curve to measure iterative quality
- Yield curve to measure ranking quality
- Cross-validation for testing with training data
- Active learning to select better training examples

Training up a classifier

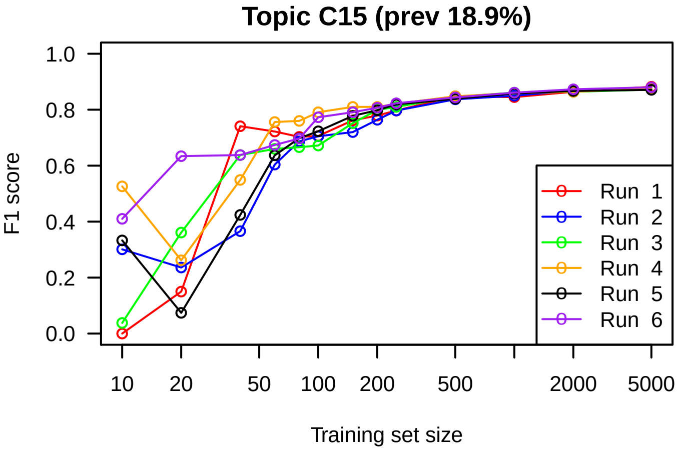
- To date, assumed all training examples available at once
- However, classifiers often trained iteratively:
  - Select, label, add training examples
  - Check classifier effectiveness
  - Repeat if not effective enough
- Training examples often require human judgment
  - Can be expensive to collect
- Only want to train as many examples as required

Learning curve



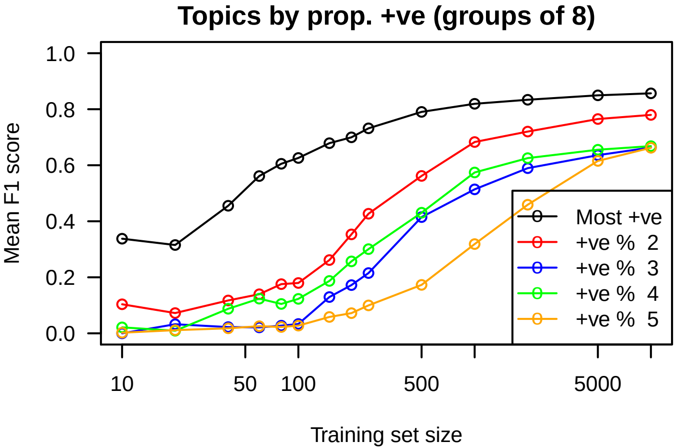
- The bigger the training set, the better the classifier
- As training examples added, classifier effectiveness improves
- But some maximum limit on effectiveness
- Due to inherent ambiguity in topic, data

Learning curve



- Different training sets lead to same plateau
- But reach there at different rates
- Would like to pick training examples to reach there faster

Variance in learning rate between topics

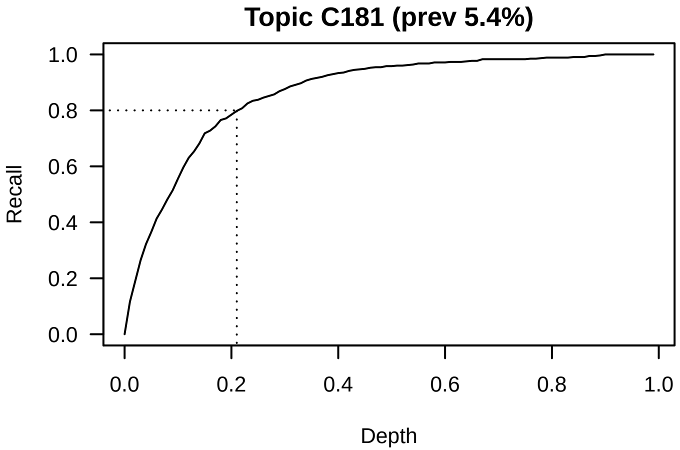


- Some topics are conceptually harder
- All other things equal, learning rate follows propotion positive:
  - The greater the proportion positive ($< 50\%$)
  - ...the faster the learning

Classification as ranking (pseudo-regression)

- Most binary classifiers can give us a strength of prediction score
- This is pseudo-regression (binary label in, real-value out)
- Quality of ranking of independent interest:
  - Binarization step can be done separately
  - Ranking may be processed
  - User may have different precision/recall tradeoffs

Yield curve

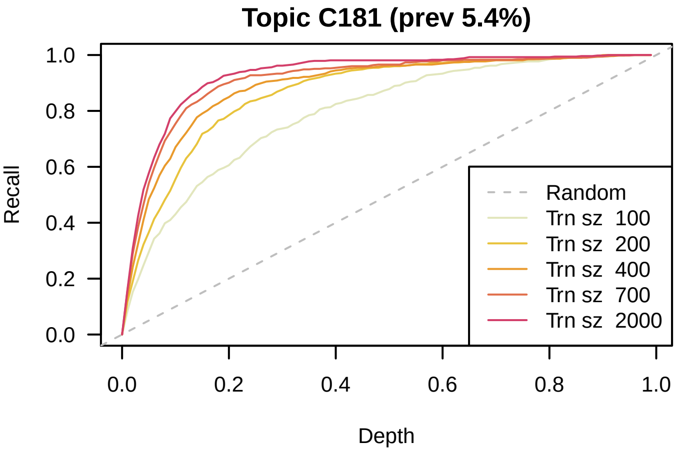


- Plotting recall against depth gives yield curve
- Indicates how far down ranking one must go to achieve give yield

Yield vs. learning curves

- NOTE: get clear in your mind difference between learning and yield curves:
  - A learning curve shows whole-classification effectiveness for increasing training sizes
  - A yield curve shows recall for different cutoff depths, for the one training size

Yield curve with increasing training



- View as yield curve, increasing training aims to push curve "up and to left"

Real-valued metrics on rankings

Ranking quality also measurable by various real-valued metrics:

- Area under curve (for whatever curve)
- Average precision
- Any other binary IR ranking metric

Testing on training

- Effectiveness experimentally measured by:
  - Training on a training set
  - Evaluating against a (separate) test set
- Testing directly on the training data exaggerates effectiveness
  - Model has been fit to training data
  - Will perform better on training data than new data
    - Though testing on training can give indication of "separability" of training data
- However, sometimes we want to reuse training set for testing:
  - We have limited labelled data
  - We are trying to tune parameters during an actual run
- One technique for reusing training data for testing is *cross-validation*

Cross-validation

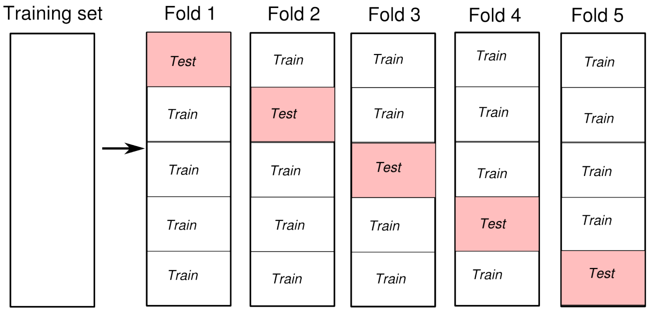


Figure : 5-fold cross-validation

- Break training set into n folds
- Successively:
  - Train on $n - 1$ folds
  - Test on n'th fold
- Aggregate scores (confusion matrices) across four folds

Limitations to cross-fold validation

- Only predicts performance on unlabelled examples if training examples a random sample from unlabelled examples
- n-fold CV predicts effectiveness of classifier with $(n-1)T/n$ training examples, not all T training examples
- Tricky to get an aggregate ranking from cross-validation
  - Because pseudo-regressed scores for different folds come from different models

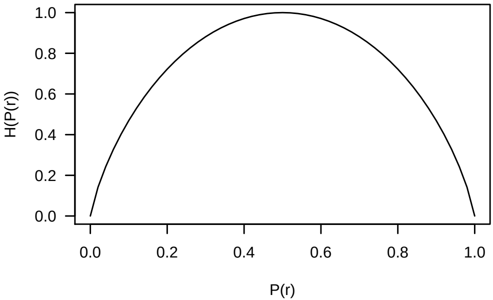
Active learning

- Some documents are better training examples than others
- Trying to select good training documents is *active learning*
- (Selecting documents at random is *passive learning*:)
- We can get the machine learner itself to help us find good training documents

Active learning by uncertainty sampling

- Ideally, like to select training documents classifier gets wrong
- Little gain in labelling training examples classifier has right
- We don't know what's wrong, right until we've labelled them
- Instead, select documents classifier is "most uncertain" about

Maximum uncertainty in probabilitic models

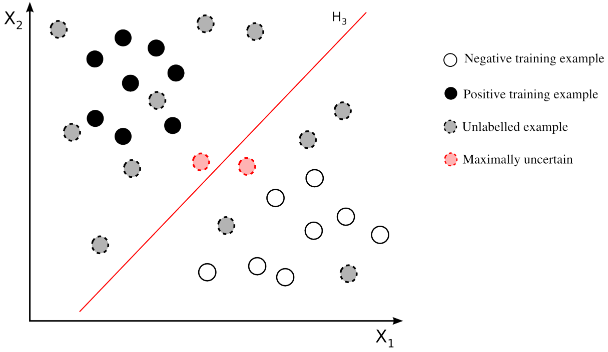


$$H(p) = -p \log_2(p) - (1-p) \log_2(1-p) \tag{1}$$

In probabilistic models (e.g. Logistic Regression)

- Most uncertain documents are those with $P(r) \approx 0.5$
- Can formalize as entropy $H(P(r))$
  - Maximized at $P(r) = 0.5$ (see figure above)

Maximum uncertainty in partitioning models



In partitioning models (e.g. SVM)

- Most uncertain are closest to separating hyper-plane
- Closest elements tend to have biggest impact on hyperplane

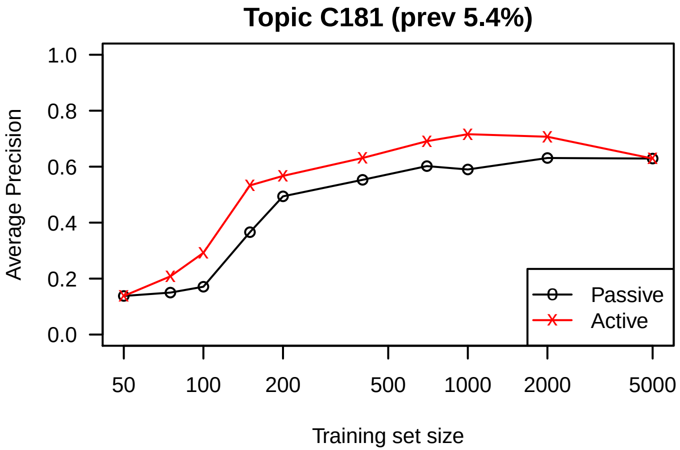
Uncertainty through CV

Another way of measuring uncertainty is through cross-validation:

- Build n models each with $(n-1)/n$ of training data
- Classify unlabelled examples with each fold-model
- Select example(s) on which fold-models most disagree

Known as "query by commitee".

Effectivenss of active learning



- Active learning typically leads to steeper learning curves
- (i.e. faster learning)
- However, there can be "degenerate cases", where active learning gets "stuck" in unproductive part of space

Active learning practicalities

- Theoretical work often assumes only one example chosen at each active iteration
- Active learning expensive
  - Must run classifier over all unlabelled examples at each iteration
  - Unlabelled examples can be very large set
  - Often inefficient to have human labeller look at only single instance at each iteration
- In practice, typically label several (perhaps tens of) examples per iteration

Selecting multiple examples

- Simple approach is to pick m most uncertain examples
  - E.g. m examples with probability of relevance closest to 50%
  - or m examples closest to separating hyper-plane
- However, examples close to given "point" in space more likely to be similar than examples further away in space
- Inefficient to label many similar examples
- Quick fix is to sample from larger set of uncertain documents

Diversifying active example selection

- Two criteria to satisfy when selecting examples:
  - Select diverse examples
  - Avoid outliers
    - Documents that are dissimilar to all others give little help
- Diversity achievable by clustering, select documents from different clusters
- Outliers avoided by outlier detection (finding documents that are far from other documents)

Looking back and forward

Back

▶

Looking back and forward

Back

- Labelling data frequently expensive
- Classifiers often iteratively trained until desired effectiveness achieved
- Progress in training measured by learning curve
- Cross validation also usable for measuring effectiveness on training data
- Binary classifiers may produce rankings
- Effectiveness of ranking measurable by yield curve
- As well as standard IR rank metrics like AP

Looking back and forward

Back

- Some training examples more useful than others
- Active learning seeks to pick most useful training examples at each iteration
- Usefulness measurable by uncertainty; either:
  - Documents closest to “decision boundary”
  - Documents which committee of (CV) classifiers disagree on
- Diversity, non-outliers important criteria for multiple selection active learning

Looking back and forward

Forward

- Topic modelling

Further reading

- Lewis and Gale, "A sequential algorithm for training text classifiers", SIGIR, 1994 (early work on active learning and uncertainty sampling)
- Xu, Akella, and Zhang, "Incorporating diversity and density in active learning for relevance feedback", ECIR, 2007 (select diverse, non-outlier examples in multiple-document active learning)
- Tong and Keller, "Support vector machine active learning with applications to text classification", JMLR 2002 (active learning techniques specific to support vector machines)
- Liere and Tadepalli, "Active learning with committees for text categorization", AAAI 1997 (query by commitee for active learning selection)