Performance of Classifiers

Performance of Classifiers

How do we evaluate whether our classifier (for documents) is any good?

Performance of Classifiers

How do we evaluate whether our classifier (for documents) is any good?

Think about a two class problem.

Performance of Classifiers

How do we evaluate whether our classifier (for documents) is any good?

Think about a two class problem. Suppose we have particular interest in identifying all the Jihadist documents.

Performance of Classifiers

How do we evaluate whether our classifier (for documents) is any good?

Think about a two class problem. Suppose we have particular interest in identifying all the Jihadist documents.

TP the document should be placed in c,

Performance of Classifiers

How do we evaluate whether our classifier (for documents) is any good?

Think about a two class problem. Suppose we have particular interest in identifying all the Jihadist documents.

TP the document should be placed in c, and method placed it in c,

Performance of Classifiers

How do we evaluate whether our classifier (for documents) is any good?

Think about a two class problem. Suppose we have particular interest in identifying all the Jihadist documents.

TP the document should be placed in c, and method placed it in c, we have a true positive.

Performance of Classifiers

How do we evaluate whether our classifier (for documents) is any good?

Think about a two class problem. Suppose we have particular interest in identifying all the Jihadist documents.

TP the document should be placed in c, and method placed it in c, we have a true positive.

FP the document should be placed in $\neg c$,

Performance of Classifiers

How do we evaluate whether our classifier (for documents) is any good?

Think about a two class problem. Suppose we have particular interest in identifying all the Jihadist documents.

- **TP** the document should be placed in c, and method placed it in c, we have a true positive.
- **FP** the document should be placed in $\neg c$, and method placed it in c,

Performance of Classifiers

How do we evaluate whether our classifier (for documents) is any good?

Think about a two class problem. Suppose we have particular interest in identifying all the Jihadist documents.

- **TP** the document should be placed in c, and method placed it in c, we have a true positive.
- **FP** the document should be placed in $\neg c$, and method placed it in c, we have a false positive (type I error).

Performance of Classifiers

How do we evaluate whether our classifier (for documents) is any good?

Think about a two class problem. Suppose we have particular interest in identifying all the Jihadist documents.

- TP the document should be placed in c, and method placed it in c, we have a true positive.
- FP the document should be placed in $\neg c$, and method placed it in c, we have a false positive (type I error).
- FN the document should be placed in c,

Performance of Classifiers

How do we evaluate whether our classifier (for documents) is any good?

Think about a two class problem. Suppose we have particular interest in identifying all the Jihadist documents.

- **TP** the document should be placed in c, and method placed it in c, we have a true positive.
- **FP** the document should be placed in $\neg c$, and method placed it in c, we have a false positive (type I error).
- **FN** the document should be placed in c, and method placed it in $\neg c$,

Performance of Classifiers

How do we evaluate whether our classifier (for documents) is any good?

Think about a two class problem. Suppose we have particular interest in identifying all the Jihadist documents.

- TP the document should be placed in c, and method placed it in c, we have a true positive.
- FP the document should be placed in $\neg c$, and method placed it in c, we have a false positive (type I error).
- FN the document should be placed in c, and method placed it in $\neg c$, we have a false negative (type II error).

Performance of Classifiers

How do we evaluate whether our classifier (for documents) is any good?

Think about a two class problem. Suppose we have particular interest in identifying all the Jihadist documents.

- **TP** the document should be placed in c, and method placed it in c, we have a true positive.
- **FP** the document should be placed in $\neg c$, and method placed it in c, we have a false positive (type I error).
- **FN** the document should be placed in c, and method placed it in $\neg c$, we have a false negative (type II error).
- **TN** the document should be placed in $\neg c$,

Performance of Classifiers

How do we evaluate whether our classifier (for documents) is any good?

Think about a two class problem. Suppose we have particular interest in identifying all the Jihadist documents.

- TP the document should be placed in c, and method placed it in c, we have a true positive.
- FP the document should be placed in $\neg c$, and method placed it in c, we have a false positive (type I error).
- FN the document should be placed in c, and method placed it in $\neg c$, we have a false negative (type II error).
- TN the document should be placed in $\neg c$, and method placed it in $\neg c$,

Performance of Classifiers

How do we evaluate whether our classifier (for documents) is any good?

Think about a two class problem. Suppose we have particular interest in identifying all the Jihadist documents.

- **TP** the document should be placed in c, and method placed it in c, we have a true positive.
- **FP** the document should be placed in $\neg c$, and method placed it in c, we have a false positive (type I error).
- **FN** the document should be placed in c, and method placed it in $\neg c$, we have a false negative (type II error).
- **TN** the document should be placed in $\neg c$, and method placed it in $\neg c$, we have a true negative.

Confusion Matrix

Confusion Matrix

		Predicted		
		J	$\neg J$	Total
Actual	J	a TP	b FN	$a+b$
	$\neg J$	c FP	d TN	$c+d$
	Total	$a+c$	$b+d$	N

Confusion Matrix

		Predicted		
		J	$\neg J$	Total
Actual	J	a TP	b FN	$a+b$
	$\neg J$	c FP	d TN	$c+d$
	Total	$a+c$	$b+d$	N

Accuracy : $\frac{\text{number correctly classified}}{\text{total number of cases}} = \frac{a+d}{a+b+c+d}$

Confusion Matrix

		Predicted		
		J	$\neg J$	Total
Actual	J	a TP	b FN	$a+b$
	$\neg J$	c FP	d TN	$c+d$
	Total	$a+c$	$b+d$	N

Accuracy : $\frac{\text{number correctly classified}}{\text{total number of cases}} = \frac{a+d}{a+b+c+d}$

Precision : $\frac{\text{number of TP}}{\text{number of TP}+\text{number of FP}} = \frac{a}{a+c}$.

Confusion Matrix

		Predicted		
		J	$\neg J$	Total
Actual	J	a TP	b FN	$a+b$
	$\neg J$	c FP	d TN	$c+d$
	Total	$a+c$	$b+d$	N

Accuracy : $\frac{\text{number correctly classified}}{\text{total number of cases}} = \frac{a+d}{a+b+c+d}$

Precision : $\frac{\text{number of TP}}{\text{number of TP}+\text{number of FP}} = \frac{a}{a+c}$.
Fraction of the documents predicted to be J, that were in fact J.

Confusion Matrix

		Predicted		
		J	$\neg J$	Total
Actual	J	a TP	b FN	$a+b$
	$\neg J$	c FP	d TN	$c+d$
	Total	$a+c$	$b+d$	N

Accuracy : $\frac{\text{number correctly classified}}{\text{total number of cases}} = \frac{a+d}{a+b+c+d}$

Precision : $\frac{\text{number of TP}}{\text{number of TP}+\text{number of FP}} = \frac{a}{a+c}$.
Fraction of the documents predicted to be J, that were in fact J.

Recall : $\frac{\text{number of TP}}{\text{number of TP} + \text{number of FN}} = \frac{a}{a+b}$.

Confusion Matrix

		Predicted		
		J	$\neg J$	Total
Actual	J	a TP	b FN	$a+b$
	$\neg J$	c FP	d TN	$c+d$
	Total	$a+c$	$b+d$	N

Accuracy : $\frac{\text{number correctly classified}}{\text{total number of cases}} = \frac{a+d}{a+b+c+d}$

Precision : $\frac{\text{number of TP}}{\text{number of TP+number of FP}} = \frac{a}{a+c}$.
Fraction of the documents predicted to be J, that were in fact J.

Recall : $\frac{\text{number of TP}}{\text{number of TP + number of FN}} = \frac{a}{a+b}$.
Fraction of the documents that were in fact J, that method predicted were J.

Confusion Matrix

		Predicted		
		J	$\neg J$	Total
Actual	J	a TP	b FN	$a+b$
	$\neg J$	c FP	d TN	$c+d$
	Total	$a+c$	$b+d$	N

Accuracy : $\frac{\text{number correctly classified}}{\text{total number of cases}} = \frac{a+d}{a+b+c+d}$

Precision : $\frac{\text{number of TP}}{\text{number of TP+number of FP}} = \frac{a}{a+c}$.
Fraction of the documents predicted to be J, that were in fact J.

Recall : $\frac{\text{number of TP}}{\text{number of TP + number of FN}} = \frac{a}{a+b}$.
Fraction of the documents that were in fact J, that method predicted were J.

F : $2\frac{\text{precision}\cdot\text{recall}}{\text{precision+recall}}$. Harmonic mean of precision and recall.

Confusion Matrix

		Predicted		
		J	$\neg J$	Total
Actual	J	a TP	b FN	$a+b$
	$\neg J$	c FP	d TN	$c+d$
	Total	$a+c$	$b+d$	N

Accuracy : $\frac{\text{number correctly classified}}{\text{total number of cases}} = \frac{a+d}{a+b+c+d}$

Precision : $\frac{\text{number of TP}}{\text{number of TP+number of FP}} = \frac{a}{a+c}$.
Fraction of the documents predicted to be J, that were in fact J.

Recall : $\frac{\text{number of TP}}{\text{number of TP + number of FN}} = \frac{a}{a+b}$.
Fraction of the documents that were in fact J, that method predicted were J.

F : $2\frac{\text{precision}\cdot\text{recall}}{\text{precision}+\text{recall}}$. Harmonic mean of precision and recall.

Exercise

Exercise

Exercise

You are working for the CIA, looking for emails that pertain to terrorist attacks.

Exercise

You are working for the CIA, looking for emails that pertain to terrorist attacks. Fortunately, such emails are very, very rare (0.0001% of all emails).

Exercise

You are working for the CIA, looking for emails that pertain to terrorist attacks. Fortunately, such emails are very, very rare (0.0001% of all emails).

1 For such a task,

Exercise

You are working for the CIA, looking for emails that pertain to terrorist attacks. Fortunately, such emails are very, very rare (0.0001% of all emails).

1 For such a task, there's probably a trade-off between precision and recall. Explain why.

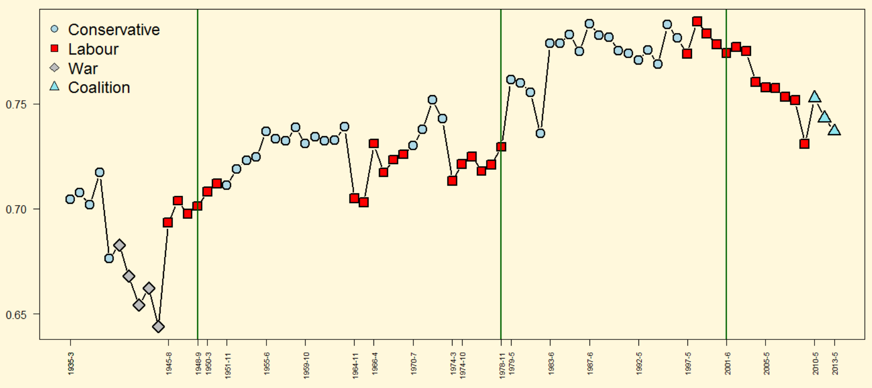
Exercise

You are working for the CIA, looking for emails that pertain to terrorist attacks. Fortunately, such emails are very, very rare (0.0001% of all emails).

1. For such a task, there's probably a trade-off between precision and recall. Explain why.
2. We may be skeptical of using accuracy as a performance indicator in this case. Explain why.

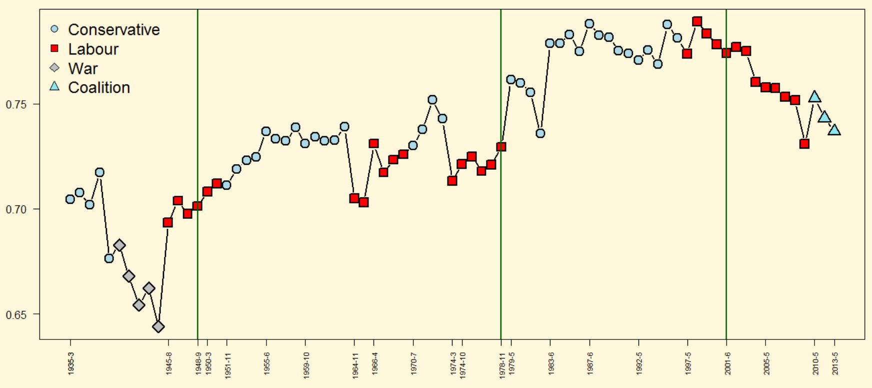
Aside: Sometimes Classifier Performance is <u>Substantively</u> Meaningful

Aside: Sometimes Classifier Performance is Substantively Meaningful



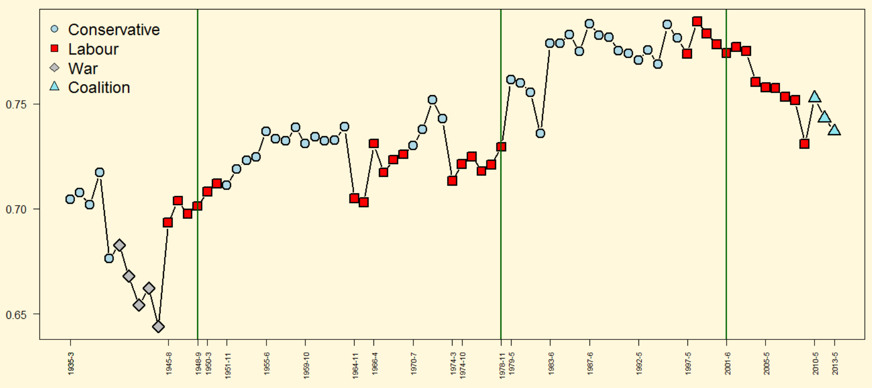
Use machine to classify left (-1) vs right ($+1$) MPs in UK and record classification accuracy.

Aside: Sometimes Classifier Performance is Substantively Meaningful



Use machine to classify left (-1) vs right ($+1$) MPs in UK and record classification accuracy. When high, parties are more polarized.

Aside: Sometimes Classifier Performance is Substantively Meaningful



Use machine to classify left (-1) vs right ($+1$) MPs in UK and record classification accuracy. When high, parties are more polarized. Makes sense in terms of historical record!

Crowdsourcing

Crowdsourcing

So far, the methods have assumed that we already have a training set,

Crowdsourcing

So far, the methods have assumed that we already have a training set, which will typically have been coded by experts.

Crowdsourcing

So far, the methods have assumed that we already have a training set, which will typically have been coded by experts.

but that can be very expensive,

Crowdsourcing

So far, the methods have assumed that we already have a training set, which will typically have been coded by experts.

but that can be very expensive, and it would be good to make it easier to replicate.

Crowdsourcing

So far, the methods have assumed that we already have a training set, which will typically have been coded by experts.

but that can be very expensive, and it would be good to make it easier to replicate.

if we had a large number of 'experts',

Crowdsourcing

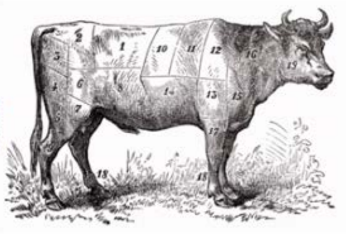
So far, the methods have assumed that we already have a training set, which will typically have been coded by experts.

but that can be very expensive, and it would be good to make it easier to replicate.

if we had a large number of 'experts', we could (depending on the size of the problem) have everything as a 'training' set and avoid modeling at all.

Galton and the Wisdom of Crowds

Galton and the Wisdom of Crowds



average of 800 guesses = 1,197
actual weight of the ox = 1,198

Crowdsourcing as Concept

Crowdsourcing as Concept

Benoit, Conway, Lauderdale, Laver and Mikhaylov (2016)

Crowdsourcing as Concept

Benoit, Conway, Lauderdale, Laver and Mikhaylov (2016) note classification jobs could be given to a large number of relatively cheap online workers.

Crowdsourcing as Concept

Benoit, Conway, Lauderdale, Laver and Mikhaylov (2016) note classification jobs could be given to a large number of relatively cheap online workers.

If those workers make the same judgements ('this document is left wing, this document is right wing') when faced with the same stimuli (on average),

Crowdsourcing as Concept

Benoit, Conway, Lauderdale, Laver and Mikhaylov (2016) note classification jobs could be given to a large number of relatively cheap online workers.

If those workers make the same judgements ('this document is left wing, this document is right wing') when faced with the same stimuli (on average), then the set of them together should obtain the truth (on average) (to the extent that is well-defined!)

Crowdsourcing as Concept

Benoit, Conway, Lauderdale, Laver and Mikhaylov (2016) note classification jobs could be given to a large number of relatively cheap online workers.

If those workers make the same judgements (‘this document is left wing, this document is right wing’) when faced with the same stimuli (on average), then the set of them together should obtain the truth (on average) (to the extent that is well-defined!)

NB Don’t care whether they are ‘representative’ of some broader population or not:

Crowdsourcing as Concept

Benoit, Conway, Lauderdale, Laver and Mikhaylov (2016) note classification jobs could be given to a large number of relatively cheap online workers.

If those workers make the same judgements ('this document is left wing, this document is right wing') when faced with the same stimuli (on average), then the set of them together should obtain the truth (on average) (to the extent that is well-defined!)

NB Don't care whether they are 'representative' of some broader population or not: this is not a survey to estimate their opinions of the labels—

Crowdsourcing as Concept

Benoit, Conway, Lauderdale, Laver and Mikhaylov (2016) note classification jobs could be given to a large number of relatively cheap online workers.

If those workers make the same judgements ('this document is left wing, this document is right wing') when faced with the same stimuli (on average), then the set of them together should obtain the truth (on average) (to the extent that is well-defined!)

NB Don't care whether they are 'representative' of some broader population or not: this is not a survey to estimate their opinions of the labels—we care about the labels themselves.

Crowdsourcing as Concept

Benoit, Conway, Lauderdale, Laver and Mikhaylov (2016) note classification jobs could be given to a large number of relatively cheap online workers.

If those workers make the same judgements ('this document is left wing, this document is right wing') when faced with the same stimuli (on average), then the set of them together should obtain the truth (on average) (to the extent that is well-defined!)

NB Don't care whether they are 'representative' of some broader population or not: this is not a survey to estimate their opinions of the labels—we care about the labels themselves.

BTW crowdsourcing can certainly be used for such 'survey' tasks—

Crowdsourcing as Concept

Benoit, Conway, Lauderdale, Laver and Mikhaylov (2016) note classification jobs could be given to a large number of relatively cheap online workers.

If those workers make the same judgements ('this document is left wing, this document is right wing') when faced with the same stimuli (on average), then the set of them together should obtain the truth (on average) (to the extent that is well-defined!)

NB Don't care whether they are 'representative' of some broader population or not: this is not a survey to estimate their opinions of the labels—we care about the labels themselves.

BTW crowdsourcing can certainly be used for such 'survey' tasks—see Berinsky et al (2012) for a review of Mechanical Turk for political science use.

Crowdsourcing in practice

Crowdsourcing in practice

BCLLM study data from the Manifesto Project:

Crowdsourcing in practice

BCLLM study data from the Manifesto Project: sentence labels by experts,

Crowdsourcing in practice

BCLLM study data from the Manifesto Project: sentence labels by experts, from over 4000 manifestos.

Crowdsourcing in practice

BCLLM study data from the Manifesto Project: sentence labels by experts, from over 4000 manifestos.

Break up manifestos into random sentences (in context),

Crowdsourcing in practice

BCLLM study data from the Manifesto Project: sentence labels by experts, from over 4000 manifestos.

Break up manifestos into random sentences (in context), and ask CrowdFlower workers to classify into economic or social policy,

Crowdsourcing in practice

BCLLM study data from the Manifesto Project: sentence labels by experts, from over 4000 manifestos.

Break up manifestos into random sentences (in context), and ask CrowdFlower workers to classify into economic or social policy, and then into one of 5 categories ('very left/liberal'–'very right/conservative').

Crowdsourcing in practice

BCLLM study data from the Manifesto Project: sentence labels by experts, from over 4000 manifestos.

Break up manifestos into random sentences (in context), and ask CrowdFlower workers to classify into economic or social policy, and then into one of 5 categories ('very left/liberal'–'very right/conservative').

Model allows for correcting for reader and text fixed effects,

Crowdsourcing in practice

BCLLM study data from the Manifesto Project: sentence labels by experts, from over 4000 manifestos.

Break up manifestos into random sentences (in context), and ask CrowdFlower workers to classify into economic or social policy, and then into one of 5 categories ('very left/liberal'–'very right/conservative').

Model allows for correcting for reader and text fixed effects, though simply taking means works well.

Crowdsourcing in practice

BCLLM study data from the Manifesto Project: sentence labels by experts, from over 4000 manifestos.

Break up manifestos into random sentences (in context), and ask CrowdFlower workers to classify into economic or social policy, and then into one of 5 categories ('very left/liberal'–'very right/conservative').

Model allows for correcting for reader and text fixed effects, though simply taking means works well.

NB can reduce uncertainty around crowd estimates by increasing number of workers for that sentence.

Crowdsourcing in practice

BCLLM study data from the Manifesto Project: sentence labels by experts, from over 4000 manifestos.

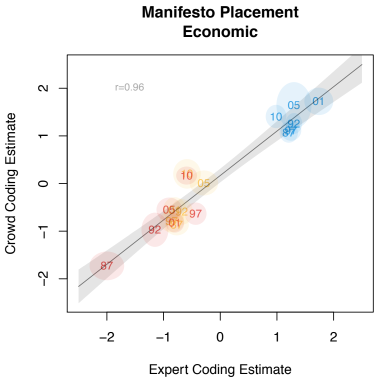
Break up manifestos into random sentences (in context), and ask CrowdFlower workers to classify into economic or social policy, and then into one of 5 categories ('very left/liberal'–'very right/conservative').

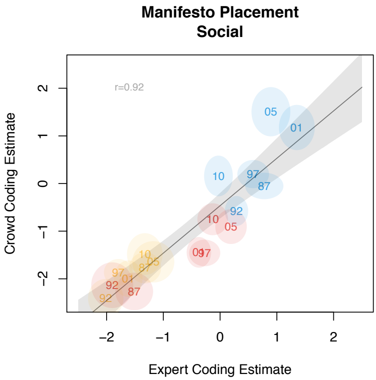
Model allows for correcting for reader and text fixed effects, though simply taking means works well.

NB can reduce uncertainty around crowd estimates by increasing number of workers for that sentence.

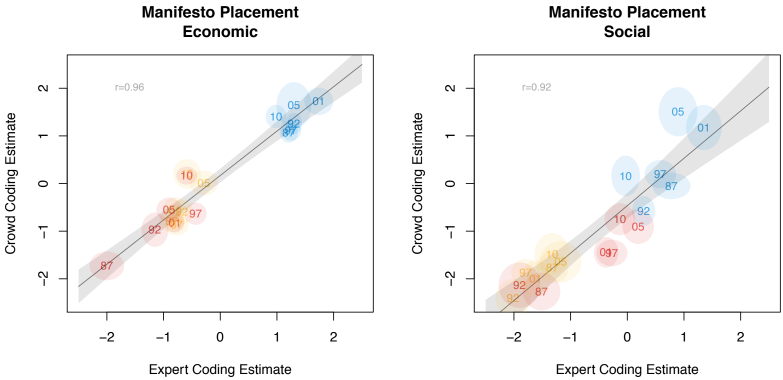
Comparing Experts and CF workers

Comparing Experts and CF workers



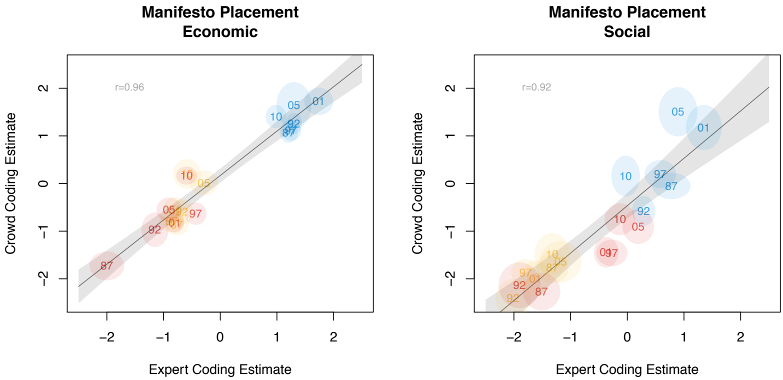


Comparing Experts and CF workers



Note that this method allows replication of the data used in an analysis,

Comparing Experts and CF workers



Note that this method allows replication of the data used in an analysis, not just the analysis itself!