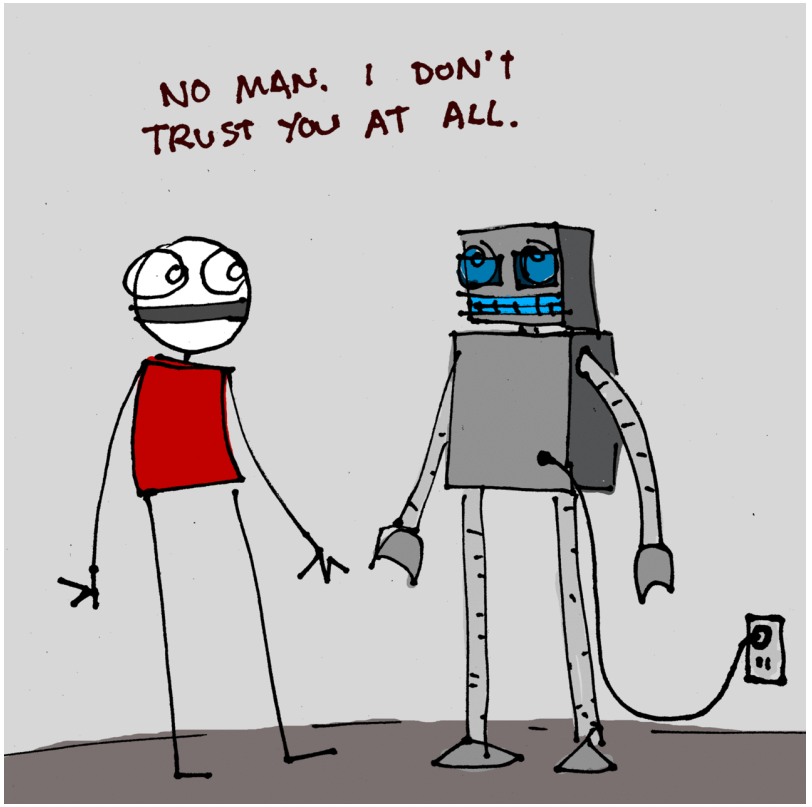
“Why Should I Trust You?” - Debugging black-box text classifiers

Tobias Sterbak / PyData Amsterdam / 27-08-18

When do you trust a model?



- metrics are not enough
- a validation set is not enougth

What is a black-box model?

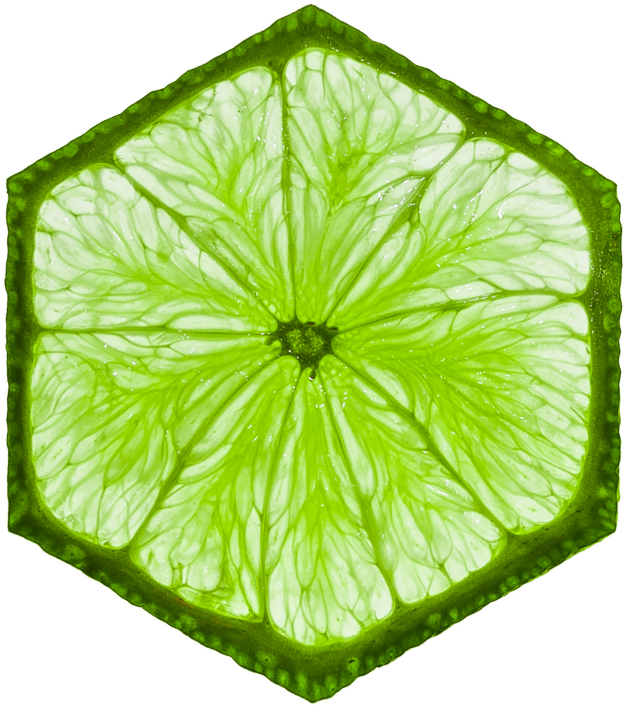
A system where the internal workings are completly hidden from you.

Examples:

- Deep neural network
- even a linear model with bag of words

Outline

- The LIME algorithm
- Example and Code
- How to make it fail



The LIME algorithm

Ribeiro et al, 2016

GOAL: understand the prediction of an arbitrary model for a certrain sample

L ocally

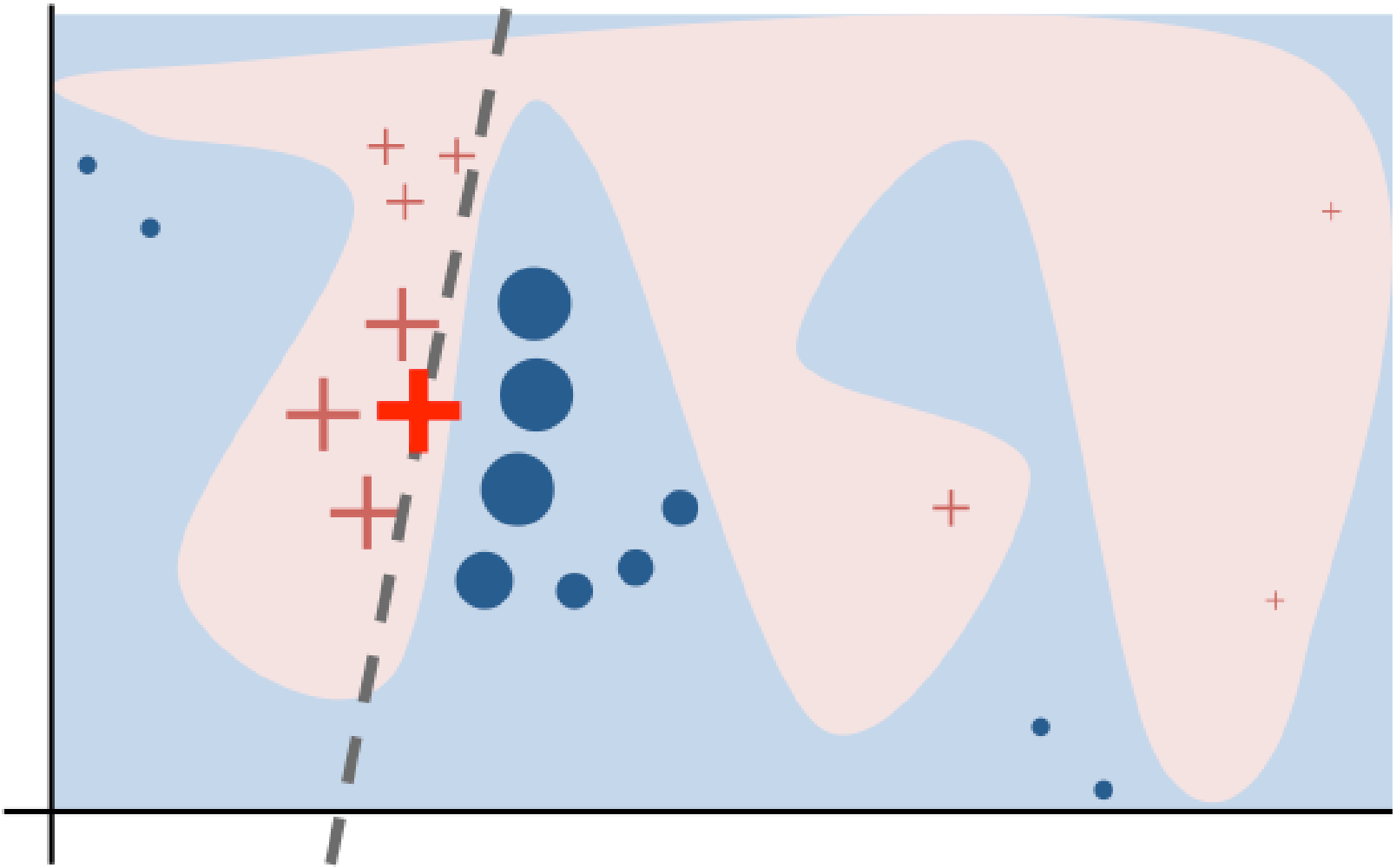
I nterpretable

M odel-agnostic

E xplanations

How it works

- Generate a fake dataset X from the example
- Use trained black-box model f to get predictions y_p for each example in a generated dataset
- Train a white-box model g on X, y_p
- Explain the original example through weights of the white-box model
- Assess how well the white-box model approximates the black-box model



Source: Ribeiro et al, 2016

Let's look at an example: the 20 newsgroup data.

In [1]:
```
from sklearn.datasets import fetch_20newsgroups
categories = ['alt.atheism', 'soc.religion.christian', 'comp.graphics', 'sci.me
d']
twenty_train = fetch_20newsgroups(subset='train', categories=categories, shuffle
=True,
                                  random_state=42, remove=('headers', 'footers'
))
twenty_test = fetch_20newsgroups(subset='test', categories=categories, shuffle=T
rue,
                                 random_state=42, remove=('headers', 'footers'))
```

In [2]:
```
i = 125
print("Class: {}".format(twenty_train.target_names[twenty_train.target[i]]))
print("-"*20); print()
sample = twenty_train.data[i]; print(sample)
```

```
Class: alt.atheism
--------------------

In article <1993Apr3.153552.4334@mac.cc.macalstr.edu>, acooper@mac.cc.macalst
r.edu writes:
|> In article <1pint5$1l4@fido.asd.sgi.com>, livesey@solntze.wpd.sgi.com (Jon
Livesey) writes
>
> Well, Germany was hardly the ONLY country to discriminate against the
> Jews, although it has the worst reputation because it did the best job
> of expressing a general European dislike of them.  This should not turn
> into a debate on antisemitism, but you should also point out that Luther's
>  antiSemitism was based on religious grounds, while Hitler's was on racial
> grounds, and Wagnmer's on aesthetic grounds.  Just blanketing the whole
> group is poor analysis, even if they all are bigots.

I find these to be intriguing remarks.   Could you give us a bit
more explanation here?   For example, which religion is anti-semitic,
and which aesthetic?
```

We train a black-box classifier.

In [25]:
```
# LSA features
vec = TfidfVectorizer(min_df=3, stop_words='english', ngram_range=(1, 2))
svd = TruncatedSVD(n_components=100, n_iter=7, random_state=42)
lsa = make_pipeline(vec, svd)

# SVM with rbf-kernel
clf = SVC(C=150, gamma=2e-2, probability=True, kernel="rbf")
text_clf = make_pipeline(lsa, clf)

text_clf.fit(twenty_train.data, twenty_train.target)
print("Accuracy: {:.1%}".format(text_clf.score(twenty_test.data, twenty_test.tar
get)))
```

```
Accuracy: 89.0%
```

Let's explain the predictions of this model

In [6]:
```
print(sample)
```

```
In article <1993Apr3.153552.4334@mac.cc.macalstr.edu>, acooper@mac.cc.macalst
r.edu writes:
|> In article <1pint5$1l4@fido.asd.sgi.com>, livesey@solntze.wpd.sgi.com (Jon
Livesey) writes
>
> Well, Germany was hardly the ONLY country to discriminate against the
> Jews, although it has the worst reputation because it did the best job
> of expressing a general European dislike of them.  This should not turn
> into a debate on antisemitism, but you should also point out that Luther's
>  antiSemitism was based on religious grounds, while Hitler's was on racial
> grounds, and Wagnmer's on aesthetic grounds.  Just blanketing the whole
> group is poor analysis, even if they all are bigots.

I find these to be intriguing remarks.   Could you give us a bit
more explanation here?   For example, which religion is anti-semitic,
and which aesthetic?
```

Look at a perturbed sample to this instance

In [38]:
```
print(get_perturbed_sample(sample))
```

```
article writes:
|> In article Livesey) Well, discriminate against the has it
> a them. not into but you out based religious grounds, while and aesthetic ev
en these For which anti-semitic,
and aesthetic?
```

Setup the explainer model

In [26]:
```
explainer = Pipeline([
    ("BoW", CountVectorizer()),                          # interpretable representa
tion
    ("selectK", SelectKBest(k=10, score_func=chi2)),  # limit the complexity of
 the explanation
    ("lr", LogisticRegression())                         # weighted interpretable m
odel
])
```

Get a lot of perturbed samples and predict on them

In [27]:
```
perturbed_samples = [get_perturbed_sample(sample) for i in range(5000)]
perturbed_predictions = text_clf.predict(perturbed_samples)
```

Fit the explainer model on the predictions of the text classifier

In [29]:
```
explainer.fit(perturbed_samples, perturbed_predictions, lr__sample_weight=weights)
```

Out[29]:
```
Pipeline(memory=None,
     steps=[('BoW', CountVectorizer(analyzer='word', binary=False, decode_error='strict',
        dtype=<class 'numpy.int64'>, encoding='utf-8', input='content',
        lowercase=True, max_df=1.0, max_features=None, min_df=1,
        ngram_range=(1, 1), preprocessor=None, stop_words=None,
        strip_...ty='l2', random_state=None, solver='liblinear', tol=0.0001,
          verbose=0, warm_start=False))])
```

How well does it approximate the black-box model?

In [30]:
```
new_pert_samples = [get_perturbed_sample(sample) for i in range(1000)]
print("Score: {:.1%}".format(explainer.score(new_pert_samples, text_clf.predict(
new_pert_samples))))
```

```
Score: 89.6%
```

Now get the important features/words

In [31]:
```
inv_vocab = {v: k for k, v in explainer.steps[0][1].vocabulary_.items()}
important_words = [inv_vocab[i] for i in explainer.steps[1][1].pvalues_.argsort
()[:10][::-1]]
```

Look at the explanation

In [32]:
```
p = explainer.predict_proba([sample])[0]
for c in explainer.steps[2][1].classes_:
    print("> {}".format(twenty_train.target_names[c]))
    print(f"probability: {p[c]:.1%}"); print("-"*20)
    for w, v in zip(important_words, explainer.steps[2][1].coef_[c]):
        print(f"{w}: {v:.3}")
    print()
```

```
> alt.atheism
probability: 100.0%
--------------------
cc: 1.07
article: 2.02
writes: 0.259
was: 2.12
grounds: 0.168
livesey: 0.211
sgi: 2.02
com: 0.348
on: 0.116
the: 1.5

> comp.graphics
probability: 0.0%
--------------------
cc: -1.05
article: -1.62
writes: -0.363
was: -1.37
grounds: 1.08
livesey: -0.22
sgi: -1.62
com: -0.341
on: -0.318
the: -1.36

> sci.med
```

In [15]:
```
print(sample)
```

```
In article <1993Apr3.153552.4334@mac.cc.macalstr.edu>, acooper@mac.cc.macalst
r.edu writes:
|> In article <1pint5$1l4@fido.asd.sgi.com>, livesey@solntze.wpd.sgi.com (Jon
Livesey) writes
>
> Well, Germany was hardly the ONLY country to discriminate against the
> Jews, although it has the worst reputation because it did the best job
> of expressing a general European dislike of them.  This should not turn
> into a debate on antisemitism, but you should also point out that Luther's
>  antiSemitism was based on religious grounds, while Hitler's was on racial
> grounds, and Wagnmer's on aesthetic grounds.  Just blanketing the whole
> group is poor analysis, even if they all are bigots.

I find these to be intriguing remarks.   Could you give us a bit
more explanation here?   For example, which religion is anti-semitic,
and which aesthetic?
```

Let's look at eli5

- python package: https://github.com/TeamHG-Memex/eli5 (https://github.com/TeamHG-Memex/eli5)
- provides insights in different model
- provides nice visualization
- allows for multiple different explainers
- kernel density estimation to get better perturbed samples

In [16]:
```
import eli5
from eli5.lime import TextExplainer

te = TextExplainer(random_state=42)
te.fit(sample, text_clf.predict_proba)
te.show_prediction(target_names=twenty_train.target_names)
```

Out[16]:

y=alt.atheism (probability **0.999**, score **8.880**) top features

Contribution?	Feature
+9.421	Highlighted in text (sum)
-0.542	<BIAS>

in article <1993apr3.153552.4334@mac.cc.macalstr.edu>, acooper@mac.cc.macalstr.edu writes: |> in article <1pint5$1l4@fido.asd.sgi.com>, livesey@solntze.wpd.sgi.com (jon livesey) writes > > well, germany was hardly the only country to discriminate against the > jews, although it has the worst reputation because it did the best job > of expressing a general european dislike of them. this should not turn > into a debate on antisemitism, but you should also point out that luther's > antisemitism was based on religious grounds, while hitler's was on racial > grounds, and wagnmer's on aesthetic grounds. just blanketing the whole > group is poor analysis, even if they all are bigots. i find these to be intriguing remarks. could you give us a bit more explanation here? for example, which religion is anti-semitic, and which aesthetic?

y=comp.graphics (probability **0.000**, score **-7.740**) top features

Contribution?	Feature
-0.016	<BIAS>
-7.725	Highlighted in text (sum)

in article <1993apr3.153552.4334@mac.cc.macalstr.edu>, acooper@mac.cc.macalstr.edu writes: |> in article <1pint5$1l4@fido.asd.sgi.com>, livesey@solntze.wpd.sgi.com (jon livesey) writes > > well, germany was hardly the only country to discriminate against the > jews, although it has the worst reputation because it did the best job > of expressing a general european dislike of them. this should not turn > into a debate on antisemitism, but you should also point out that luther's > antisemitism was based on religious grounds, while hitler's was on racial > grounds, and wagnmer's on aesthetic grounds. just blanketing the whole > group is poor analysis, even if they all are bigots. i find these to be intriguing remarks. could you give us a bit more explanation here? for example, which religion is anti-semitic, and which aesthetic?

y=sci.med (probability **0.000**, score **-12.524**) top features

Contribution?	Feature
-0.196	<BIAS>
-12.328	Highlighted in text (sum)

in article <1993apr3.153552.4334@mac.cc.macalstr.edu>, acooper@mac.cc.macalstr.edu writes: |> in article <1pint5$1l4@fido.asd.sgi.com>, livesey@solntze.wpd.sgi.com (jon livesey) writes > > well, germany was hardly the only

country to discriminate against the > jews, although it has the worst reputation because it did the best job > of expressing a general european dislike of them. this should not turn > into a debate on antisemitism, but you should also point out that luther's > antisemitism was based on religious grounds, while hitler's was on racial > grounds, and wagnmer's on aesthetic grounds. just blanketing the whole > group is poor analysis, even if they all are bigots. i find these to be intriguing remarks. could you give us a bit more explanation here? for example, which religion is anti-semitic, and which aesthetic?

y=soc.religion.christian (probability **0.000**, score **-9.564**) top features

Contribution?	Feature
-0.215	<BIAS>
-9.348	Highlighted in text (sum)

in article <1993apr3.153552.4334@mac.cc.macalstr.edu>, acooper@mac.cc.macalstr.edu writes: |> in article <1pint5$1l4@fido.asd.sgi.com>, livesey@solntze.wpd.sgi.com (jon livesey) writes > > well, germany was hardly the only country to discriminate against the > jews, although it has the worst reputation because it did the best job > of expressing a general european dislike of them. this should not turn > into a debate on antisemitism, but you should also point out that luther's > antisemitism was based on religious grounds, while hitler's was on racial > grounds, and wagnmer's on aesthetic grounds. just blanketing the whole > group is poor analysis, even if they all are bigots. i find these to be intriguing remarks. could you give us a bit more explanation here? for example, which religion is anti-semitic, and which aesthetic?

How to trick the algorithm

In [17]:
```
def predict_proba_len(docs):
    proba = [
        [0, 1.0, 0.0, 0] if len(doc) % 2 else [1.0, 0, 0, 0]
        for doc in docs
    ]
    return np.array(proba)
```

In [18]:
```
len(sample)
```

Out[18]:
```
850
```

In [19]:
```
te2 = TextExplainer().fit(sample, predict_proba_len)
te2.show_prediction(target_names=twenty_train.target_names)
```

Out[19]: **y=comp.graphics** (probability **0.546**, score **0.183**) top features

Contribution?	Feature
+0.255	Highlighted in text (sum)
-0.072	<BIAS>

in article <1993apr3.153552.4334@mac.cc.macalstr.edu>, acooper@mac.cc.macalstr.edu writes: |> in article <1pint5$1l4@fido.asd.sgi.com>, livesey@solntze.wpd.sgi.com (jon livesey) writes > > well, germany was hardly the only country to discriminate against the > jews, although it has the worst reputation because it did the best job > of expressing a general european dislike of them. this should not turn > into a debate on antisemitism, but you should also point out that luther's > antisemitism was based on religious grounds, while hitler's was on racial > grounds, and wagnmer's on aesthetic grounds. just blanketing the whole > group is poor analysis, even if they all are bigots. i find these to be intriguing remarks. could you give us a bit more explanation here? for example, which religion is anti-semitic, and which aesthetic?

We can detect this failure by **looking at** metrics:

In [20]:
```
te2.metrics_
```

Out[20]:
```
{'mean_KL_divergence': 0.72973696077940187, 'score': 0.48996566001842468}
```

Luckily it's possible to fix this.

In [22]:
```
class DocLength(TransformerMixin):
    def fit(self, X, y=None):
        return self

    def transform(self, X):
        return [[len(doc) % 2, not len(doc) % 2] for doc in X]

    def get_feature_names(self):
        return ['is_odd', 'is_even']

vec = make_union(DocLength(), CountVectorizer(ngram_range=(1,2)))
te3 = TextExplainer(vec=vec).fit(sample, predict_proba_len)
```

In [23]:
```
print(te3.metrics_)
te3.explain_prediction(target_names=twenty_train.target_names)
```

```
{'mean_KL_divergence': 0.011840784438819842, 'score': 1.0}
```

Out[23]:

y=alt.atheism (probability **0.988**, score **-4.448**) top features

Contribution?	Feature
+4.419	doclength__is_even
+0.015	countvectorizer: Highlighted in text (sum)
+0.014	<BIAS>

countvectorizer: in article <1993apr3.153552.4334@mac.cc.macalstr.edu>, acooper@mac.cc.macalstr.edu writes: |> in article <1pint5$1l4@fido.asd.sgi.com>, livesey@solntze.wpd.sgi.com (jon livesey) writes > > well, germany was hardly the only country to discriminate against the > jews, although it has the worst reputation because it did the best job > of expressing a general european dislike of them. this should not turn > into a debate on antisemitism, but you should also point out that luther's > antisemitism was based on religious grounds, while hitler's was on racial > grounds, and wagnmer's on aesthetic grounds. just blanketing the whole > group is poor analysis, even if they all are bigots. i find these to be intriguing remarks. could you give us a bit more explanation here? for example, which religion is anti-semitic, and which aesthetic?

Tl;dl

- Inspect your models not only by looking at validation metrics
- LIME can help you to get some understanding of your model (and eli5 makes it easy)
- It's important to understand the “lenses” you’re looking through when using LIME
- Never trust an alogrithm blindly!

Where to find me

www.depends-on-the-definition.com

www.github.com/tsterbak

@tobias_sterbak

Questions?