THE COST OF CALIBRATION

Kilian Q. Weinberger, Cornell University
Doc, I am not feeling well. I have symptoms X, Y, Z.

Let me feed your data into my classifier to see if you have disease Q.
In Deep-Nets calibration is easily achievable.

[Guo et al., ICML 2017]

Calibration has unexpected and undesirable side-effects.

[Kleinberg et al., NIPS 2017]
CALIBRATED PROBABILITY ESTIMATES

Input → Classifier → Prediction

well calibrated probability estimate

Conf. = 0.7
CALIBRATED PROBABILITY ESTIMATES

Input ➔ Classifier ➔ Prediction

well calibrated probability estimate

Conf. = 0.7
\( \mathbb{E} [y] = 0.7 \)
VISUALIZING MISCALIBRATION

Bin A

0.5 0.5 0.5 0.5 0.5
0.5 0.5 0.5 0.5 0.5

Conf. = 0.5

Bin B

0.7 0.7 0.7 0.7 0.7
0.7 0.7 0.7 0.7 0.7

Conf. = 0.7

Bin C

0.9 0.9 0.9 0.9 0.9
0.9 0.9 0.9 0.9 0.9

Conf. = 0.9
VISUALIZING MISCALIBRATION

<table>
<thead>
<tr>
<th>Bin A</th>
<th>Bin B</th>
<th>Bin C</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>0.7</td>
<td>0.9</td>
</tr>
<tr>
<td>0.5</td>
<td>0.7</td>
<td>0.9</td>
</tr>
<tr>
<td>0.5</td>
<td>0.7</td>
<td>0.9</td>
</tr>
<tr>
<td>0.5</td>
<td>0.7</td>
<td>0.9</td>
</tr>
<tr>
<td>0.5</td>
<td>0.7</td>
<td>0.9</td>
</tr>
</tbody>
</table>

Conf. = 0.5, 0.7, 0.9

\[ \mathbb{E}[y|A] = 0.4 \]
\[ \mathbb{E}[y|B] = 0.7 \]
\[ \mathbb{E}[y|C] = 0.8 \]

Gap: 0.1, 0.0, 0.1
VISUALIZING MISCALIBRATION

Graph showing confidence levels and output.

- Bin A: Conf. = 0.5, E[y | A] = 0.4
- Bin B: Conf. = 0.7, E[y | B] = 0.7
- Bin C: Conf. = 0.9, E[y | C] = 0.8

Confidence

Gap:

0.1
0.0
0.1
Predicting Good Probabilities With Supervised Learning

Alexandru Niculescu-Mizil
Rich Caruana
Department Of Computer Science, Cornell University, Ithaca NY 14853

Abstract
We examine the relationship between the predictions made by different learning algorithms and true posterior probabilities. We show that maximum margin methods such as boosted trees and boosted stumps push probability mass away from 0 and 1 yielding a characteristic sigmoid shaped distortion in the predicted probabilities. Models such as Naïve Bayes, which make unrealistic independence assumptions, push probabilities toward 0 and 1. Other models such as neural nets and bagged trees do not have these biases and predict well-calibrated probabilities. We experiment with two ways of correcting the biased probabilities predicted by some learning methods: Platt Scaling and Isotonic Regression. We qualitatively examine what kinds of distortions these calibration methods are suitable for and quantitatively examine how much data they need to be effective. The empirical results show that after calibration boosted trees, random forests, and SVMs predict the best probabilities.

1. Introduction
In many applications it is important to predict well-calibrated probabilities; good accuracy or area under the ROC curve are not sufficient. This paper examines the work such as bagged trees and neural nets have little or no bias and predict well-calibrated probabilities.

After examining the distortion (or lack of) characteristic to each learning method, we experiment with two calibration methods for correcting these distortions.


Platt Scaling is most effective when the distortion in the predicted probabilities is sigmoid-shaped. Isotonic Regression is a more powerful calibration method that can correct any monotonic distortion. Unfortunately, this extra power comes at a price. A learning curve analysis shows that Isotonic Regression is more prone to overfitting, and thus performs worse than Platt Scaling, when data is scarce.

Finally, we examine how good are the probabilities predicted by each learning method after each method’s predictions have been calibrated. Experiments with eight classification problems suggest that random forests, neural nets and bagged decision trees are the best learning methods for predicting well-calibrated probabilities prior to calibration, but after calibration the best methods are boosted trees, random forests and SVMs.

2. Calibration Methods
MISCALIBRATION DEEPNETS

LeNet (1998)

Outputs
Gap

Error=44.9

Confidence

DenseNet-40 (2017)

Outputs
Gap

Error=30.6

Confidence

Uncal. - SST-Binary
Tree LSTM

Uncal. - SST-Binary
Tree LSTM

Uncal. - CIFAR-100
DenseNet-40

Confidence

E[y]

CIFAR-100

[Guo et al., ICML 2017]
0/1 Error
DenseNet-40, CIFAR-100

[Epoch 0: ⊿ 0]
[Epoch 2: ⊿ 2]
[Epoch 4: ⊿ 4]
[Epoch 6: ⊿ 6]
[Epoch 8: ⊿ 8]

Error

Training
Test

Epoch

[Gao et al., ICML 2017]
DenseNet-40, CIFAR-100

NLL

Epoch

Train

Test

[Gao et al., ICML 2017]
TEMPERATURE SCALING

-5.2 0.1 2.1

[Jaynes, 1957] [Hinton et al., 2015]

$e^z(x) \over \sum_i e^{z_i(x)}$

Important to note here:
- This is a post processing step: once the rest of the network has been trained
- We optimize $T$ to minimize NLL on a holdout validation set

[Platt, 1999] [Gao et al., ICML 2017]
TEMPERATURE SCALING

[Jaynes, 1957]
[Hinton et al., 2015]

Important to note here:
- This is a post processing step: once the rest of the network has been trained
- We optimize T to minimize NLL on a holdout validation set

\[
p = \frac{e^{z(x)}}{\sum_i e^{z_i(x)}}
\]

Last Layer

Softmax

prob. simplex

class 1
\( p = 0 \)

class 2
\( p = 0.91 \)

class 3
\( p = 0.09 \)

[Platt, 1999] [Gao et al., ICML 2017]
TEMPERATURE SCALING

[Jaynes, 1957]  
[Hinton et al., 2015]

$\mathbb{P}(y) \propto e^{\frac{z(x)}{T}}$  

Important notes here:  
Temp scaling is just averaging with the uniform distribution.  
Because of this, it preserves the mode. In other words, our most probable class prediction stays the same.

[Platt, 1999]  [Gao et al., ICML 2017]
TEMPERATURE SCALING ON CIFAR-100

Uncal. - CIFAR-100
DenseNet-40

Temp. Scale - CIFAR-100
DenseNet-40

Uncal. - SST-Binary
Tree LSTM

Temp. Scale - SST-Binary
Tree LSTM
SIDE EFFECTS OF CALIBRATION
CALIBRATION

Conf. = 0.5

\[ \mathbb{E}[y \mid A] = 0.5 \]

Conf. = 0.7

\[ \mathbb{E}[y \mid B] = 0.7 \]

Conf. = 0.9

\[ \mathbb{E}[y \mid C] = 0.9 \]
CALIBRATION

\[
\mathbb{E}[\text{Conf.}] = \mathbb{E}[y]
\]

\[
\mathbb{E}[\text{Conf.}] = \mathbb{E}[\text{Conf.} \mid y = 0] \mathbb{P}[y = 0] + \mathbb{E}[\text{Conf.} \mid y = 1] \mathbb{P}[y = 1]
\]
CALIBRATION

Bin A
Conf. = 0.5
E[y|A] = 0.5

Bin B
Conf. = 0.7
E[y|B] = 0.7

Bin C
Conf. = 0.9
E[y|C] = 0.9

\[ E[y] = E[\text{Conf.} \mid y = 0] P[y = 0] + E[\text{Conf.} \mid y = 1] P[y = 1] \]

constant
constant
constant

false positive rate (FPR)

1 - (false negative rate) (FNR)
LINEARITY OF CALIBRATED CLASSIFIERS

\[ \mathbb{E}[y] = (\text{FPR}) (1 - \mathbb{E}[y]) + (1 - \text{FNR}) \mathbb{E}[y] \]
GROUP-WISE CALIBRATION

Conf. = 0.7
GROUP-WISE CALIBRATION
GROUP-WISE CALIBRATION

\[ E_{\text{women}}[\text{Conf.}] = E_{\text{women}}[y] \neq E_{\text{men}}[y] = E_{\text{men}}[\text{Conf.}] \]
GROUP-WISE CALIBRATION

Conf. = 0.6

\[ \mathbb{E}_{\text{women}} [\text{Conf.}] = \mathbb{E}_{\text{women}} [y] \neq \mathbb{E}_{\text{men}} [y] = \mathbb{E}_{\text{men}} [\text{Conf.}] \]
GROUP-WISE CALIBRATION

\[ \mathbb{E}_{\text{women}} [\text{Conf.}] = \mathbb{E}_{\text{women}} [y] \neq \mathbb{E}_{\text{men}} [y] = \mathbb{E}_{\text{men}} [\text{Conf.}] \]
ERROR DISPARITY

False Positive Rate

False Negative Rate

men

women

FNR < FNR 
women men

FPR > FPR 
women men
ERROR DISPARITY

\[ FNR_{\text{women}} = FNR_{\text{men}} \]

\[ FPR_{\text{women}} = FPR_{\text{men}} \]
Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

[https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing]
**Impossibility Theorem:**

[Kleinberg, Mullainathan, Raghavan 2016]

- **Groupwise calib.**
  - $\text{FNR}_{\text{women}} = \text{FNR}_{\text{men}}$
  - $\text{FPR}_{\text{women}} = \text{FPR}_{\text{men}}$

There will always be some disparity between two groups, if base rates aren’t equal.

Can only be mutually satisfied in trivial cases.
RELAXED CONDITIONS

Groupwise calib.

- $FNR_{\text{women}} = FNR_{\text{men}}$
- $FPR_{\text{women}} = FPR_{\text{men}}$
RELAXED CONDITIONS

Groupwise calib.

\[ \text{FNR}_{\text{women}} = \text{FNR}_{\text{men}} \]

\[ \text{FPR}_{\text{women}} = \text{FPR}_{\text{men}} \]
RELAXED CONDITIONS

Groupwise calib.

\[
\left( \alpha_w \text{FPR} + \beta_w \text{FNR} \right)_{\text{women}} = \left( \alpha_m \text{FPR} + \beta_m \text{FNR} \right)_{\text{men}}
\]
AN OPTIMAL ALGORITHM :-(

Assume we want equal FPR
The dotted line is the best we can possibly do.

“best possible” classifiers
AN OPTIMAL ALGORITHM :-(

"best possible" non-discriminatory classifiers

[ Kleinberg et al. NIPS, 2017 ]
EXAMPLE: CRIMINAL RECIDIVISM PREDICTION

Calib. + Equal F.P.

Orig. $G_1$
Orig. $G_2$
Derived $G_2$

False Neg. Rate

False Pos. Rate

COMPAS
CONCLUSION

- Calibration is important!
- Deep Nets are badly calibrated (overconfident)
- Simple temperature scaling fixes the calibration gap
- Group-wise calibration is inherently at odds with some notions of fairness
ACKNOWLEDGEMENTS

Geoff Pleiss  
Chuan Guo  
Yu Sun  
Manish Raghavan  
Felix Wu  
Jon Kleinberg
WE ARE HIRING

- Tenure-track Faculty
- Postdoc Positions
- http://cs.cornell.edu