The model is simple, until proven otherwise – how to cope in an ever changing world

Anita Faul

Laboratory for Scientific Computing, University of Cambridge

Outline

Outline:

- Motivation.
- Bayesian Learning.
- Neural Networks.
- Combination.

BBC | Anita Faul | News | Sport | Weather | iPlayer | Sounds

NEWS

Home | UK | World | Business | Politics | Tech | Science | Health | Family & Education

Science & Environment

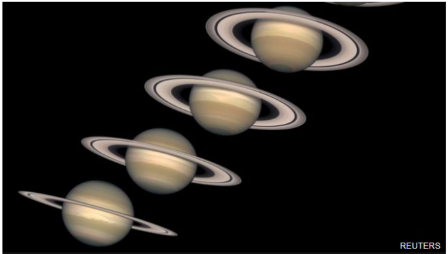
AAAS: Machine learning 'causing science crisis'

By Pallab Ghosh
Science correspondent, BBC News, Washington

16 February 2019

Share

American Association for the Advancement of Science Meeting

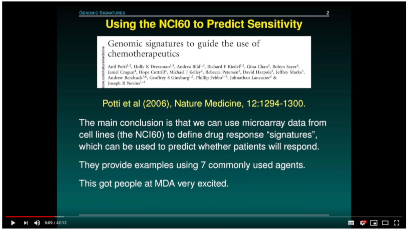


Astronomy is one of the many areas of science in which machine learning is used to make discoveries

Machine-learning techniques used by thousands of scientists to analyse data are producing results that are misleading and often completely wrong.

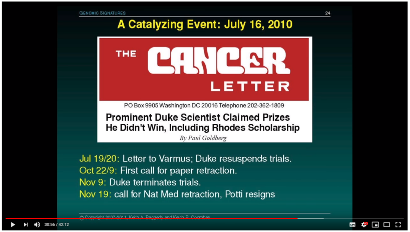
Inspiration

Keith Baggerly: Forensic Bioinformatics



The Brad Efron Honorary Symposium on LARGE-SCALE INFERENCE - Keith Baggerly

2006



The Brad Efron Honorary Symposium on LARGE-SCALE INFERENCE - Keith Baggerly

2010

Playing with people's lives.

Confusion matrix

- N: number of negative samples, P: number of positive samples.
- *True negatives* TN: number of negative samples correctly classified.
- *False positives* FP: number of negative samples misclassified.
- *True positives* TP: number of positive samples correctly classified.
- *False negatives* FN: number of positive samples misclassified.
- *Confusion table*:

	N	P
classified negative	TN	FN
classified positive	FP	TP

Sensitivity and Specificity

- *Sensitivity*, aka *true positive rate*, *recall* and *probability of detection*: fraction of positive samples correctly classified: $\text{TPR} = \frac{\text{TP}}{\text{P}}$.
- *Specificity* or *true negative rate*: fraction of negative samples correctly identified: $\text{TNR} = \frac{\text{TN}}{\text{N}}$.
- *False negative rate* or *miss rate*: $\text{FNR} = \frac{\text{FN}}{\text{P}} = 1 - \text{TPR}$.
- *False positive rate* or *fall-out*: $\text{FPR} = \frac{\text{FP}}{\text{N}} = 1 - \text{TNR}$.

Likelihood Ratios

- *Likelihood ratio for positive results*: how much more likely is positive classification in positive samples compared to in negative samples:

$$\mathrm{LR}+ = \frac{\mathrm{TPR}}{\mathrm{FPR}}.$$

- *Likelihood ratio for negative results*: how much more likely is negative classification in positive samples compared to in negative samples:

$$\mathrm{LR}- = \frac{\mathrm{FNR}}{\mathrm{TNR}}.$$

- All so far independent of prevalence.

Interpretation

- A perfect classifier would be 100% sensitive (all positives are correctly identified) and 100% specific (no negatives are incorrectly classified).
- LR$+ \to \infty$ and LR$- \to 0$.
- LR$+ > 10$ and LR$- < 0.1$ make a useful classifier according to Jaeschke R, Guyatt G, Lijmer J, *'Diagnostic tests'* in Guyatt G, Rennie D, eds. *'Users guides to the medical literature'* Chicago: AMA Press, (2002).

The New York Times

Opinion

OP-ED CONTRIBUTOR

When an Algorithm Helps Send You to Prison

By Ellora Thadaney Israni

Oct. 26, 2017

In 2013, police officers in Wisconsin arrested a man driving a car that had been used in a recent shooting. The man, Eric Loomis, pleaded guilty to attempting to flee an officer, and no contest to operating a vehicle without the owner's consent. Neither of his crimes mandates prison time.

At Mr. Loomis's sentencing, the judge cited, among other factors, Mr. Loomis's high risk of recidivism as predicted by a computer program called COMPAS, a risk assessment algorithm used by the state of Wisconsin. The judge denied probation and prescribed an 11-year sentence: six years in prison, plus five years of extended supervision.

No one knows exactly how COMPAS works; its manufacturer refuses to disclose the proprietary algorithm. We only know the final risk assessment score it spits out, which judges may consider at sentencing.

Forbes | Billionaires | Innovation | Leadership

Management AI: Bias, Criminal Recidivism, And The Promise Of Machine Learning

David A. Teich Contributor
Tirias Research Contributor Group
B2B technology analyst and consultant

Criminal recidivism is when a released criminal goes back to crime. From charging crimes through probation, the criminal justice system is constantly looking for ways to better predict which criminals are more likely to remain legal on release and who is a risk of recidivism. Bias can create inaccuracies through weighing variables incorrectly, and machine learning might provide a way of limiting bias and improving recidivism predictions.

Shutterstock

A recent study by Julia Dressel and Hany Farid, published in Science Advances, points to the limitations of deterministic algorithms with fixed parameters for the task of such predictions. The study analyzes the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) software, a package used by court systems to predict the likelihood of recidivism in

Pro Publica Inc.

Larson J, Mattu S, Kirchner L and Angwin J (2016) at Pro Publica Inc. obtained data on the re-offending risks as returned by the COMPAS algorithm and the actual occurrences of re-offending within two years after release.

	N	P	
low risk	2681	1216	3897
high risk	1282	2035	3317
	3963	3251	7214

sensitivity: $\text{TPR} = 0.63$ $\quad \text{FNR} = 0.37$ $\quad \color{red}{\text{LR}+ = 1.97}$

specificity: $\text{TNR} = 0.68$ $\quad \text{FPR} = 0.32$ $\quad \color{red}{\text{LR}- = 0.54}$

Pro Publica Inc.

Black

	N	P	
low risk	990	532	1522
high risk	805	1369	2174
	1795	1901	3696

$\text{TPR} = 0.72$

$\text{TNR} = 0.55$ $\quad$ $\text{FNR} = 0.28$

$\text{LR}+ = 1.60$ $\quad$ $\text{FPR} = 0.45$

$\text{LR}- = 0.51$

White

	N	P	
low risk	1139	461	1600
high risk	349	505	854
	1488	966	2454

$\text{TPR} = 0.52$

$\text{TNR} = 0.77$ $\quad$ $\text{FNR} = 0.48$

$\text{LR}+ = 2.26$ $\quad$ $\text{FPR} = 0.23$

$\text{LR}- = 0.62$

Challenges:

- Modeling data, keeping models simple while explaining the data adequately.
- New data arriving.
- Confidence in model predictions.
- Choice of model space.

Data Prediction problem

The data prediction problem

- We make the assumption that the data are a result of an underlying process which we do not know.
- Given measurements $t_1, \ldots, t_D$, each measurement depends on parameters we know $\mathbf{x}_1, \ldots, \mathbf{x}_D$.
- D is the dimension of the data space.
- These are quantities which can be measured with more or less effort.

Unknowns

- The measurements also depend on parameters we do not know.
- A real world application depends on factors which cannot be measured (or these measurements would be disproportionally difficult).
- For example the physics of waves are well understood. However, they depend on the medium the wave travels in, the material and its properties. These are the unknown parameters of the process.

Dictionaries

Dictionaries

- If we had a set of candidate functions $d_1(\mathbf{x}), \ldots, d_M(\mathbf{x})$, which all are solutions to the process for different parameters, we could try which fits the measurements and thus infer the underlying structure.
- We say the functions $d_1(\mathbf{x}), \ldots, d_M(\mathbf{x})$ form a dictionary and assume

$$f(\mathbf{x}) = \sum_{m=1}^{M} c_m d_m(\mathbf{x}),$$

  where $c_1, \ldots, c_M$ are coefficients and these need to be determined.
- The basis functions of the dictionary are the building blocks which build a model for the data.
- M is the dimension of the model space.

Noise

- The relationship to the measurements is

$$t_i = f(\mathbf{x}_i) + \epsilon_i.$$

- ϵ_i is noise intrinsic to the measurement process and assumed to be independent and identically, normally distributed, $\mathcal{N}(0, \sigma^2)$.

Mathematical model

$$t_i = f(\mathbf{x}_i) + \epsilon_i = \sum_{m=1}^{M} c_m d_m(\mathbf{x}_i) + \epsilon_i,$$

- Let $\mathbf{D}$ be the matrix with entries $\mathbf{D}_{i,m} = d_m(\mathbf{x}_i)$ and let $\mathbf{t}^T = (t_1, \ldots, t_D)$, $\mathbf{c}^T = (c_1, \ldots, c_M)$ and $\boldsymbol{\epsilon}^T = (\epsilon_1, \ldots, \epsilon_D)$, then

$$\mathbf{t} = \mathbf{D}\mathbf{c} + \boldsymbol{\epsilon}.$$

- $\mathbf{D}$ is an $D \times M$ matrix. However, D and M are not static. D varies with the number of measurements of the same process, while M varies with the dictionary of basis functions.

Sparse Bayesian Learning

- Central idea is that the coefficients $\mathbf{c}$ follow a distribution.
- Each coefficient c_m is a priori normally distributed with mean zero and variance α_m^{-1}.
- α_m is the precision of the distribution.
- If the precision is very large, the distribution becomes peaked at its mean and we have more confidence in the value of c_m than if it is small and the width of the distribution large.

- Multivariate prior distribution:

$$p(\mathbf{c}|\boldsymbol{\alpha}) = (2\pi)^{-M/2}\sqrt{|A|}\exp\left(\mathbf{c}^T A\mathbf{c}\right),$$

where A is a diagonal matrix with entries $A_{mm} = \alpha_m$

- Multivariate posterior distribution is normal with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$ given by

$$\boldsymbol{\Sigma} = \left(A + \sigma^{-2}D^T D\right)^{-1} \qquad \boldsymbol{\mu} = \sigma^{-2}\Sigma D^T\mathbf{t}.$$

- Since $\mathbf{t} = \mathbf{Dc} + \boldsymbol{\epsilon}$, the data is viewed as being drawn from a normal distribution with mean $\mathbf{D}\boldsymbol{\mu}$ and variance $\sigma^2\mathbf{I} + \mathbf{D\Sigma D}$.

Marginal Likelihood

- The marginal likelihood is the probability of the data given the model specified by $\mathbf{D}, \boldsymbol{\alpha}$ and σ^2 after integrating out the coefficients $\mathbf{c}$.

$$
\begin{aligned}
\mathcal{L}(\mathbf{t}|\mathbf{D}, \boldsymbol{\alpha}, \sigma^2) &= (2\pi)^{-N/2} |\sigma^2 \mathbf{I} + \sum_{m=1}^{M} \frac{1}{\alpha_m} \mathbf{d}_m \mathbf{d}_m^T|^{-1/2} \\
&\quad \exp\left(-\frac{1}{2} \mathbf{t}^T (\sigma^2 \mathbf{I} + \sum_{m=1}^{M} \frac{1}{\alpha_m} \mathbf{d}_m \mathbf{d}_m^T)^{-1} \mathbf{t} \right).
\end{aligned}
$$

- We maximize the likelihood.

Maximizing the Marginal Likelihood

- If the derivative is

positive,	we move towards a maximum,
negative,	we move away from the maximum,
zero,	we are at the maximum.

- Defining

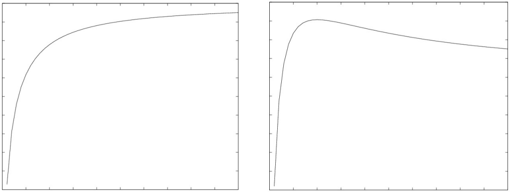
$$\mathbf{C} = \sigma^2\mathbf{I} + \sum_{m=1}^{M} \frac{1}{\alpha_m}\mathbf{d}_m\mathbf{d}_m^T, \quad \mathbf{C}_{-i} = \mathbf{C} - \frac{1}{\alpha_i}\mathbf{d}_i\mathbf{d}_i^T,$$

$$s_i = \mathbf{d}_i^T\mathbf{C}_{-i}^{-1}\mathbf{d}_i, \qquad q_i = \mathbf{d}_i^T\mathbf{C}_{-i}^{-1}\mathbf{t}.$$

- The derivative with respect to α_i of the logarithm of the marginal likelihood is

$$\underbrace{\frac{1}{2}(\alpha_i + s_i)^{-2}}_{>0}(s_i - q_i^2 + \underbrace{\frac{s_i^2}{\alpha_i}}_{\geq 0}).$$

Sparse Bayesian Learning



$s_i - q_i^2 > 0$ $\qquad$ $s_i - q_i^2 \leq 0$

In practice many α_m become infinite during maximization, meaning that the posterior distribution of the corresponding c_m is infinitely peaked at 0 and the corresponding building block can be removed from the model.

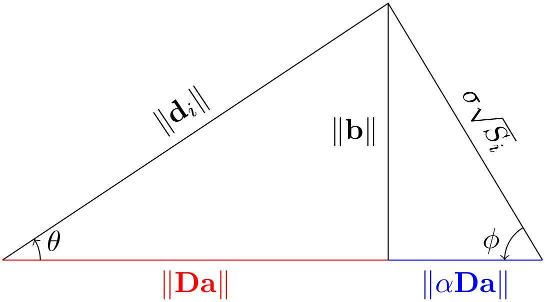
Sparsity and Quality Factor

- Let $S_i = \mathbf{d}_i^T \mathbf{C}^{-1} \mathbf{d}_i$, then $s_i = \dfrac{\alpha_i S_i}{\alpha_i - S_i}$.
- Let $Q_i = \mathbf{d}_i^T \mathbf{C}^{-1} \mathbf{t}$, then $q_i = \dfrac{\alpha_i Q_i}{\alpha_i - S_i}$.
- It can be shown that

$$\mathbf{C}^{-1}\mathbf{t} = \sigma^{-2}\left(\mathbf{t} - \mathbf{D}\boldsymbol{\mu}\right).$$

- Q_i quantifies how well aligned the building block is with this error.
- If it is orthogonal, than $\mathbf{d}_i$ will not help in removing this error.

Sparsity and Quality Factor



- By the law of sines $\dfrac{\sin\theta}{\sin\phi} = \dfrac{\sigma\sqrt{S_i}}{\|\mathbf{d}_i\|} = \sigma\sqrt{\dfrac{\mathbf{d}_i^T}{\|\mathbf{d}_i\|}\mathbf{C}^{-1}\dfrac{\mathbf{d}_i}{\|\mathbf{d}_i\|}}$.
- Si measures, how different the building block is to the others.

Model Generation

Model Generation

- Initializing the model with a single building block.
- All other hyper-parameters are notionally infinity.
- The basis function d_m, where setting its hyper-parameter α_m to its optimal value (given the current model) gives the largest increase in the marginal likelihood, is found and the model updated accordingly.
- Note that the optimal value of α_m can be finite or infinite.
  - Addition: If d_m is not in the model and the optimal α_m is finite.
  - Deletion: If d_m is in the model and the optimal α_m is infinite.
  - Re-estimation: If d_m is in the model and the optimal α_m is finite.
- This addresses the first challenge.

Predictions

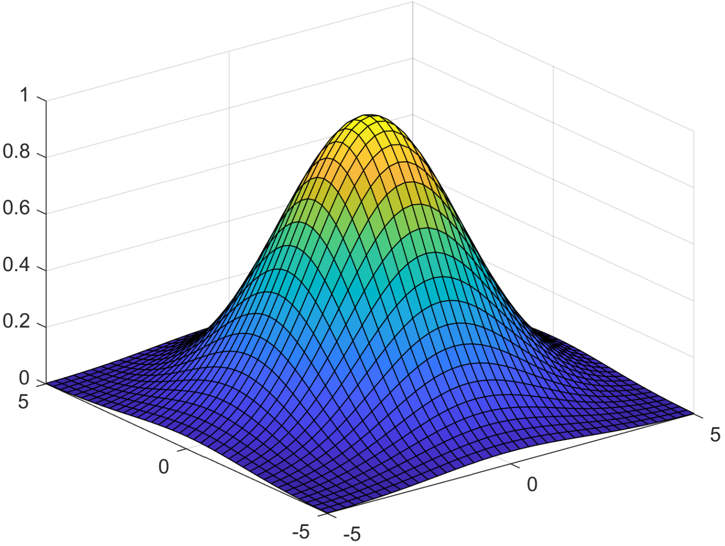
Predictions

- For a new $\mathbf{x}_*$, the predictions $t_* = \mathbf{c}^T \mathbf{d}_*$, where $\mathbf{d}_*^T = (d_1(\mathbf{x}_*), \ldots, d_M(\mathbf{x}_*))$, follow a univariate normal distribution with

$$\begin{aligned} &\text{mean} && m_* = \boldsymbol{\mu}^T \mathbf{d}_*, \\ &\text{variance} && \sigma_*^2 = \mathbf{d}_*^T \boldsymbol{\Sigma} \mathbf{d}_*, \end{aligned}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the mean and variance of the posterior distribution of the coefficients.

Predictions



New data

- A new measurement $(\mathbf{x}_*, t_*)$ means adding a row to D

$$D_* = \left(\frac{D}{\mathbf{d}_*^T} \right).$$

- The logarithm of the marginal likelihood $\log \mathcal{L}(\mathbf{t}_* | \boldsymbol{\alpha}, \sigma^2)$ is $\log \mathcal{L}(\mathbf{t} | \boldsymbol{\alpha}, \sigma^2) + \Delta \mathcal{L}$, where

$$\Delta \mathcal{L} = \log \frac{1}{\sqrt{2\pi}\sigma_*} \exp \left(- \frac{(m_* - t_*)^2}{2\sigma_*^2} \right).$$

- Thus the change is the logarithm of the likelihood of the new measurement t_* at $\mathbf{x}_*$ given the model.

New data

- Since $\sigma_* \geq \sigma$, the change lies between $-\infty$ and $\log \frac{1}{\sqrt{2\pi}\sigma}$.
- If it is positive, the new measurement affirms the model.
- If it is negative, the model is not good enough and should be updated.
- This can be done following the previous method.
- This addresses the second challenge.

Confidence in predictions

- Note, that the predictive distribution depends on the choice of basis functions. In particular, if $\mathbf{d}_* = 0$, then the mean m_* is zero, while the variance $\sigma_*^2 = \sigma^2$. The model fails completely.
- Let $\mathcal{S}$ be a subset of the samples. This could be all samples or a suitable set of neighbours of $\mathbf{x}_*$.
- We estimate the probability distribution of t_* to be normal with mean and variance

$$\begin{aligned} \bar{m} &= \operatorname*{mean}_{\mathbf{x}_i \in \mathcal{S}} \{t_i\}, \\ \bar{\sigma} &= \operatorname*{var}_{\mathbf{x}_i \in \mathcal{S}} \{t_i\}. \end{aligned}$$

Confidence in predictions

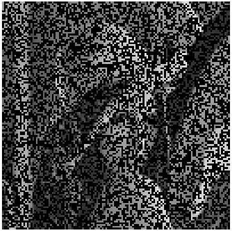
- The expected change in the logarithm of the marginal likelihood is estimated as

$$E[\Delta\mathcal{L}] = \log \frac{1}{\sqrt{2\pi}\sigma_*} - \frac{\bar{\sigma}^2+(\bar{m}-m_*)^2}{2\sigma_*^2}.$$

- If the predictive probability distribution agrees well with the estimated distribution, the change is positive and we have confidence in our predictions.
- If it does not match well, the expected change is negative, indicating that here more data should be gathered.
- This addresses the third challenge.

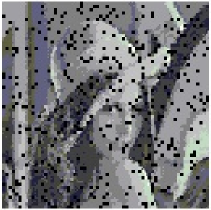
Experiments

Original

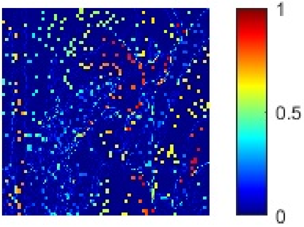


Decimation by 55%

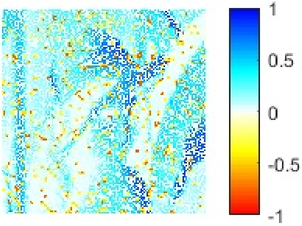
Experiments



FSIM $= 0.74$



Scaled absolute difference



Scaled $E[\Delta\mathcal{L}]$

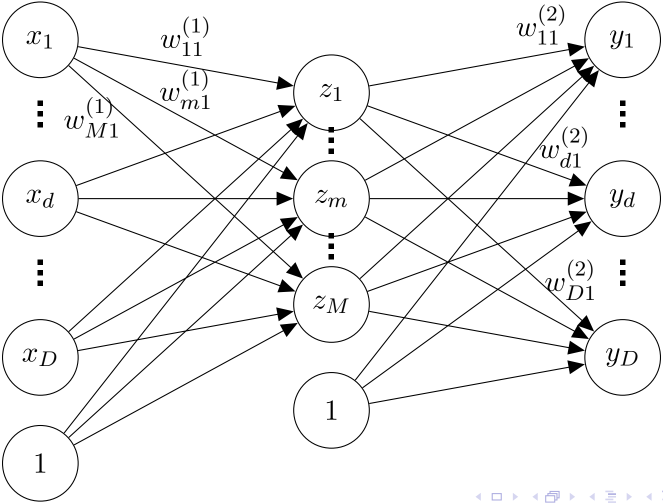
5% more samples
as informed by $E[\Delta\mathcal{L}]$
FSIM $= 0.93$

Improvements with different
basis functions
FSIM $= 0.91$

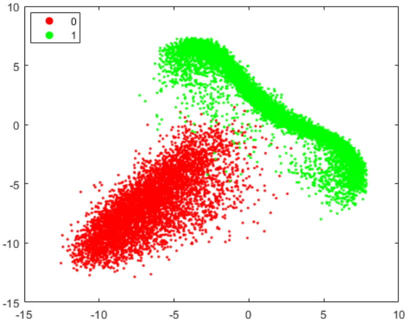
MNIST

$60,000$ images of handwritten digits of size $28 \times 28 = 784$

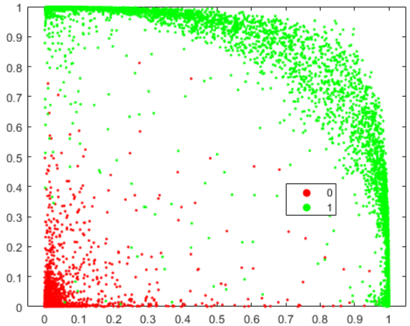
Autoencoder



2 Hidden Neurons, 2 Digits

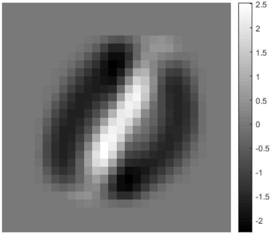


Spatial separation of activations.

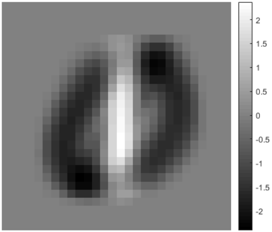


Spatial separation of latent variables.

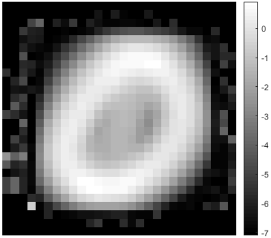
2 Hidden Neurons, 2 Digits



First basis.



Second basis.



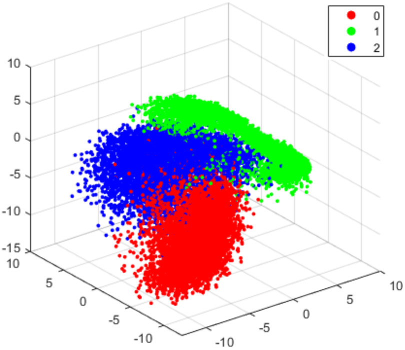
Bias.

2 Hidden Neurons, 2 Digits

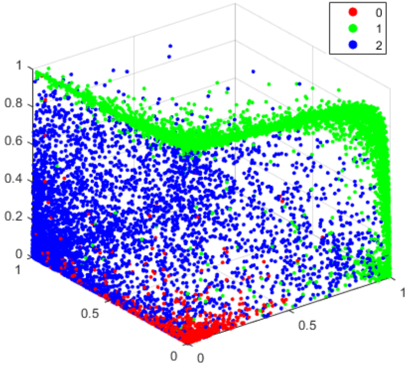
Original.

Reconstruction.

3 Hidden Neurons, 3 Digits

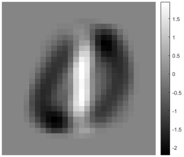


Spatial separation of activations.

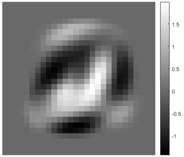


Spatial separation of latent variables.

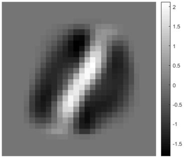
3 Hidden Neurons, 3 Digits



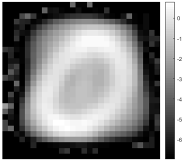
First basis.



Second basis.

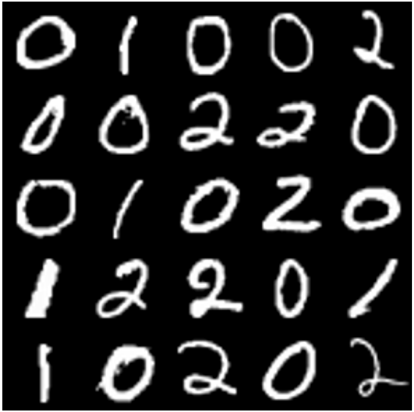


Third basis.



Bias.

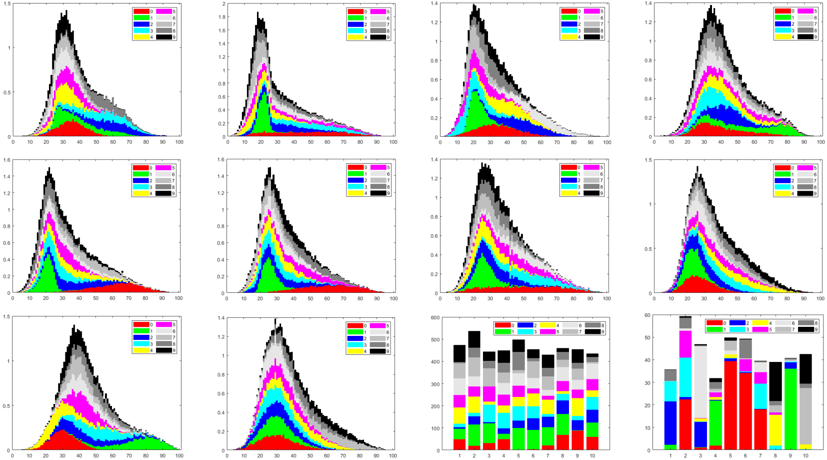
3 Hidden Neurons, 3 Digits



Original.

Reconstruction.

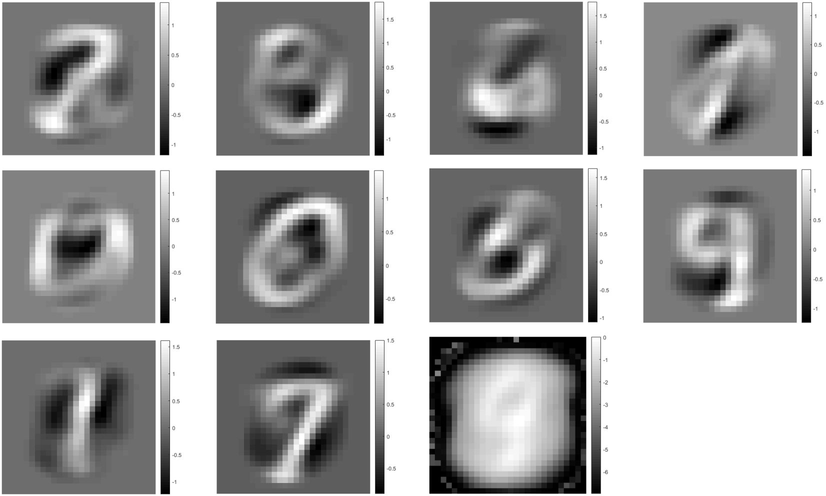
10 Hidden Neurons, 10 Digits



10 Hidden Neurons, 10 Digits

Neuron \ Digit	0	1	2	3	4	5	6	7	8	9
1			×	×					×	
2	×			×		×			×	
3			×				×			
4		×							×	
5	×							×		
6	×					×			×	
7	×			×		×	×			
8					×					×
9		×	×							
10								×		×

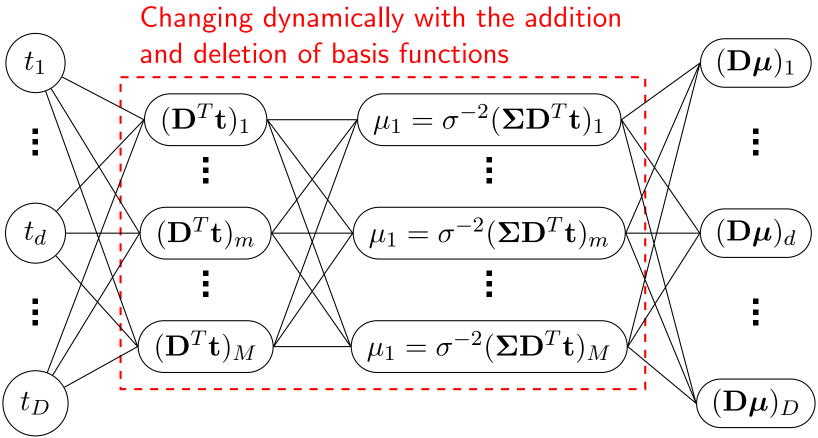
10 Hidden Neurons, 10 Digits



10 Digits, 10 Hidden Neurons

Ten digits, ten hidden neurons. ☹

$\Rightarrow$ more hidden neurons 😃


Changing dynamically with the addition and deletion of basis functions
t_1
t_d
t_D
$(\mathbf{D}^T\mathbf{t})_1$
$(\mathbf{D}^T\mathbf{t})_m$
$(\mathbf{D}^T\mathbf{t})_M$
$\mu_1 = \sigma^{-2}(\mathbf{\Sigma}\mathbf{D}^T\mathbf{t})_1$
$\mu_1 = \sigma^{-2}(\mathbf{\Sigma}\mathbf{D}^T\mathbf{t})_m$
$\mu_1 = \sigma^{-2}(\mathbf{\Sigma}\mathbf{D}^T\mathbf{t})_M$
$(\mathbf{D}\boldsymbol{\mu})_1$
$(\mathbf{D}\boldsymbol{\mu})_d$
$(\mathbf{D}\boldsymbol{\mu})_D$

Conclusions

Conclusions

- Flexible framework.
- Giving probablilistic meaning to the relevance of the model components.
- Capable to generate and include new dictionaries if new insights are gained.
- Capable to incorportate new data.
- Confidence measure for predictions.
- Guidance for the data gathering process.

Contact

Contact

- LinkedIn: https://www.linkedin.com/in/anita-faul-123750104/
- Forthcoming book: https://www.amazon.co.uk/Concise-Introduction-Machine-Learning/dp/0815384106