Reliable Probabilistic Prediction for Medical Decision Support

Harris Papadopoulos

Frederick Research Center, Cyprus

Frederick University, Cyprus

Motivation

- Most medical diagnostic systems produce bare predictions
  - They predict the most likely diagnosis without providing any information about how much one can rely on their prediction
- However, some indication about the likelihood of each diagnosis is of paramount importance in a medical setting
- Venn Predictors make multiprobability predictions
  - In effect they define lower and upper bounds for the conditional probability of the new example belonging to each class
  - These bounds are guaranteed (up to statistical fluctuations) to contain the corresponding true conditional probabilities
  - Their only assumption is that the data are i.i.d.

The Venn Prediction Framework

- We are given a set of classified examples

$$\{z_1, \ldots, z_l\} = \{(x_1, y_1), \ldots, (x_l, y_l)\}$$

  and a new unclassified example x_{l+1}.
- We want to predict the probability of x_{l+1} belonging to each class $Y_j \in \{Y_1, \ldots, Y_c\}$.
- Venn Prediction:
  - Divide all examples into a number of categories λ
  - Calculate the probability of x_{l+1} belonging to each class $Y_j \in \{Y_1, \ldots, Y_c\}$ as the frequency of Y_j in the category that contains it

Venn Taxonomy

- A *Venn Taxonomy* defines how examples are divided into categories based on some conventional ML algorithm – in this case a Neural Network.
- Assigns each example (x_i, y_i), $i = 1, \ldots, l+1$ to a category based on the output of the conventional algorithm for x_i after being trained on

$$\{(x_1,y_1),\ldots, (x_{l+1},y_{l+1})\},$$

or on

$$\{(x_1,y_1),\ldots, (x_{i-1},y_{i-1}), (x_{i+1},y_{i+1}),\ldots, (x_{l+1},y_{l+1})\}.$$

Multiprobability Predictions

- As we do not know y_{l+1}, we assume each possible class $Y_k \in \{Y_1, \ldots, Y_c\}$ for x_{l+1} in turn and partition the set

$$\{(x_1, y_1), \ldots, (x_l, y_l), (x_{l+1}, Y_k)\}$$

  into categories using a Venn taxonomy.

- For each assumed class Y_k and corresponding partitioning, we find the category T_{new} that contains x_{l+1} and calculate the empirical probability of each label Y_j in this category

$$p^{Y_k}(Y_j) = \frac{|\{(x^*, y^*) \in T_{new} : y^* = Y_j\}|}{|T_{new}|}. \qquad (1)$$

- This gives a probability distribution for each assumed class

Multiprobability Predictions

- The set of probability distributions obtained for all possible classes $Y_k \in \{Y_1, \ldots, Y_c\}$ of x_{l+1} compose the multiprobability prediction of the Venn Predictor for x_{l+1}

$$P_{l+1} = \{p^{Y_k} : Y_k \in \{Y_1, \ldots, Y_c\}\}.$$

- The probability interval for x_{l+1} belonging to Y_j is

$$\left[\min_{k=1,\ldots,c} p^{Y_k}(Y_j), \max_{k=1,\ldots,c} p^{Y_k}(Y_j)\right].$$

- If we denote this interval as $[L(Y_j), U(Y_j)]$, the prediction of the Venn Predictor is $\hat{y}_{l+1} = Y_{j_{best}}$ where

$$j_{best} = \arg\max_{j=1,\ldots,c} L(Y_j).$$

Neural Networks Venn Predictor

- Here we are interested in binary classification problems
- The NNs have one output: a value in the interval [0,1]
- Venn Taxonomy:
  - Partition the range of the NN output into λ equally sized regions
  - Each of these regions defines one category
  - Examples with output in the same region are assigned to the same category

The NN-VP Algorithm

- For every possible label $k \in \{0, 1\}$ do
  - Train the NN on the set $\{(x_1,y_1),\ldots, (x_l,y_l), (x_{l+1},k)\}$
  - Supply the input patterns $x_1,\ldots, x_{l+1}$ to the trained NN to obtain the output values $o_1,\ldots, o_{l+1}$
  - Find the category $T_i \in \{T_1,\ldots,T_\lambda\}$ that contains (x_{l+1},k)
  - Find all examples with NN output between $(i\text{-}1)/\lambda$ and i/λ and assign them to category T_i
  - Compute $p^k(1)$ and $p^k(0)$ using equation (1)
- Compute $L(0) = \min_{k=0,1} p^k(0)$ and $L(1) = \min_{k=0,1} p^k(1)$
- **Output:** the prediction $\hat{y} = \arg\max_{j=0,1} L(j)$

  and the probability interval for $\hat{y}$: $\left[\min_{k=0,1} p^k(\hat{y}), \max_{k=0,1} p^k(\hat{y})\right]$

Experiments

- Two medical datasets
  - *Mammographic Mass*: discrimination between benign and malignant mammographic masses. 961 cases: 516 benign and 445 malignant.
  - *Pima Indians Diabetes*: forecasting the onset of diabetes mellitus in a high-risk population of Pima Indians. 768 cases: 500 positive and 268 negative.
- The NNs were trained with the scaled conjugate gradient algorithm minimizing cross-entropy error
  - Their outputs can be interpreted as probabilities
- Comparison between original NN and NN-VP

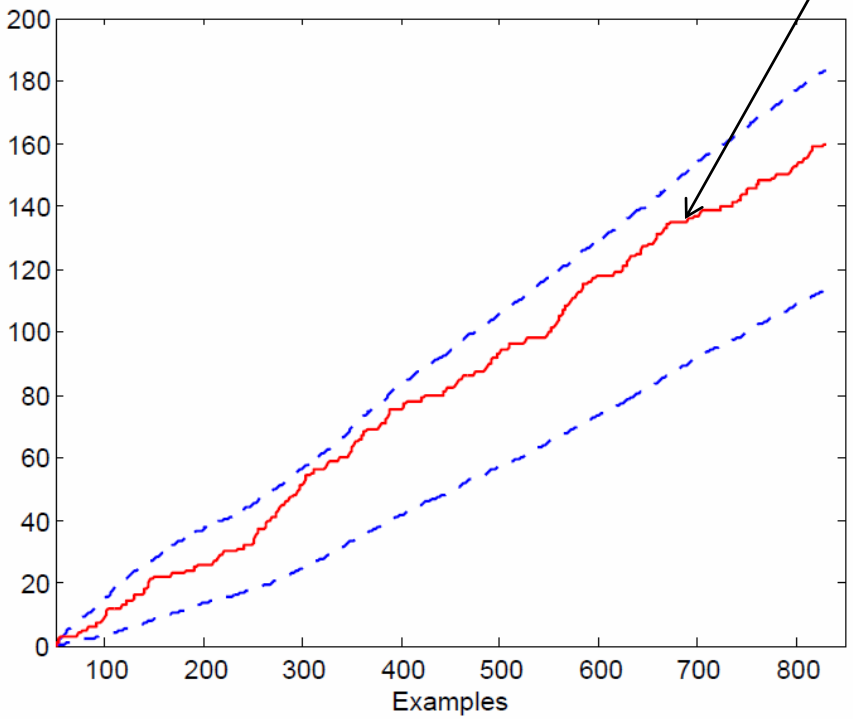
On-line Experiments

- Start with an initial training set containing 50 examples
- Predict each subsequent example
- After prediction each new example is added to the training set (with its true classification) for predicting the next examples

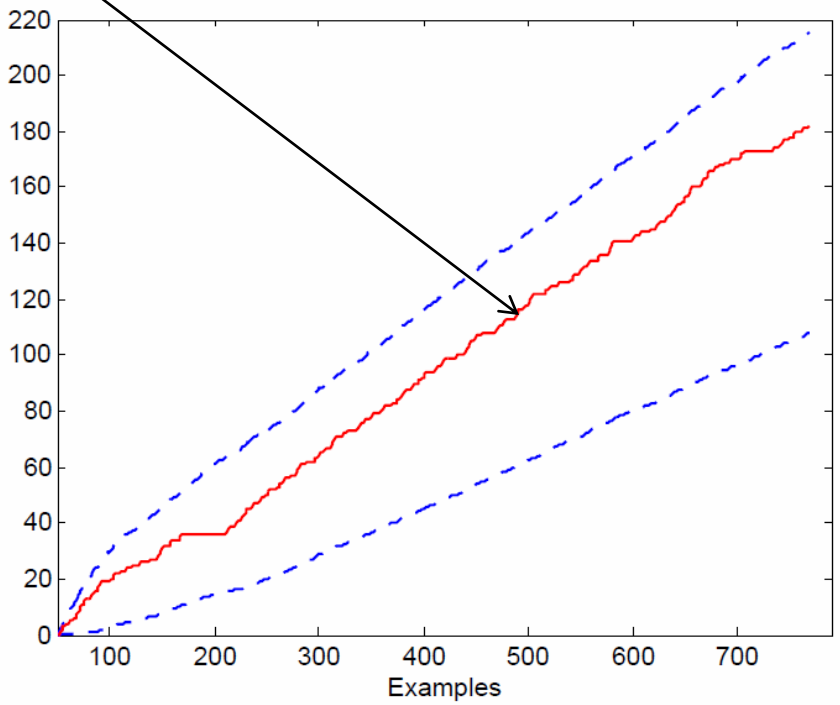
On-line Experiments: NN-VP

Cumulative Error Curve

$$E_n = \sum_{i=1}^{n} err_i$$



(a) Mammographic Mass

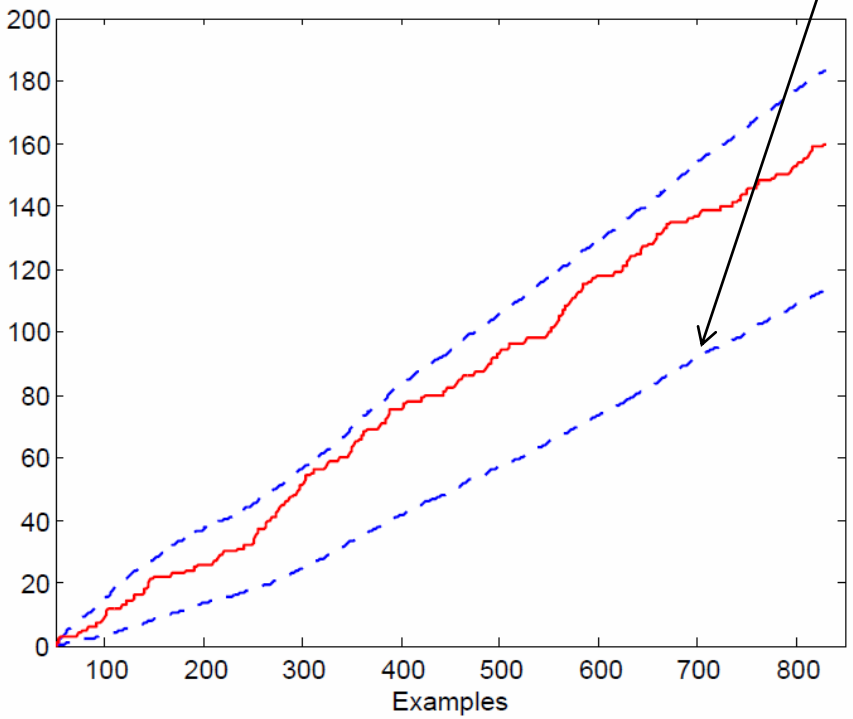


(b) Pima Indians Diabetes

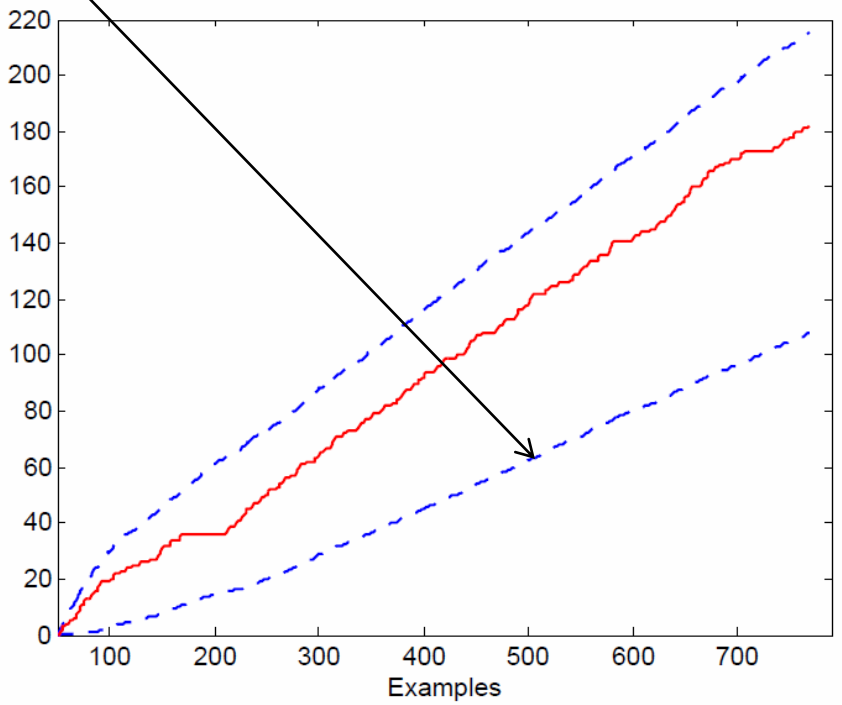
On-line Experiments: NN-VP

Cumulative Lower Error Probability Curve

$$LEP_n = \sum_{i=1}^{n} 1 - U(\hat{y}_i)$$



(a) Mammographic Mass

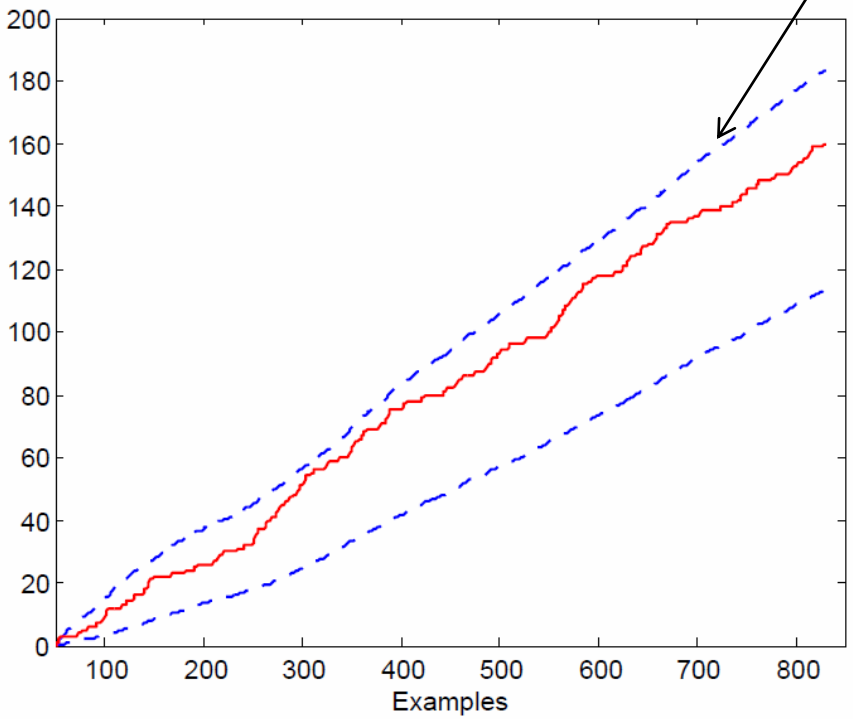


(b) Pima Indians Diabetes

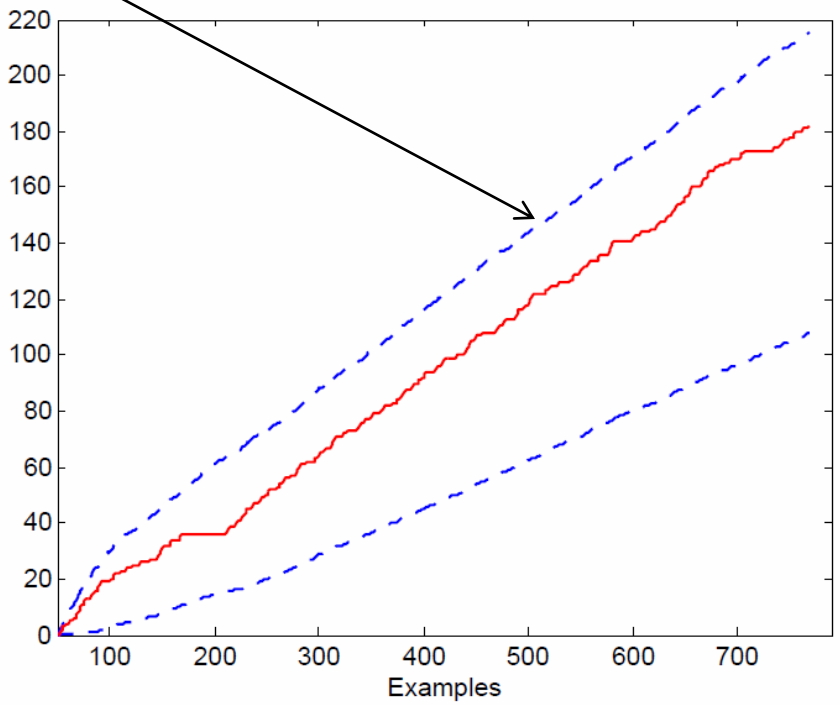
On-line Experiments: NN-VP

Cumulative Upper Error Probability Curve

$$UEP_n = \sum_{i=1}^{n} 1 - L(\hat{y}_i)$$



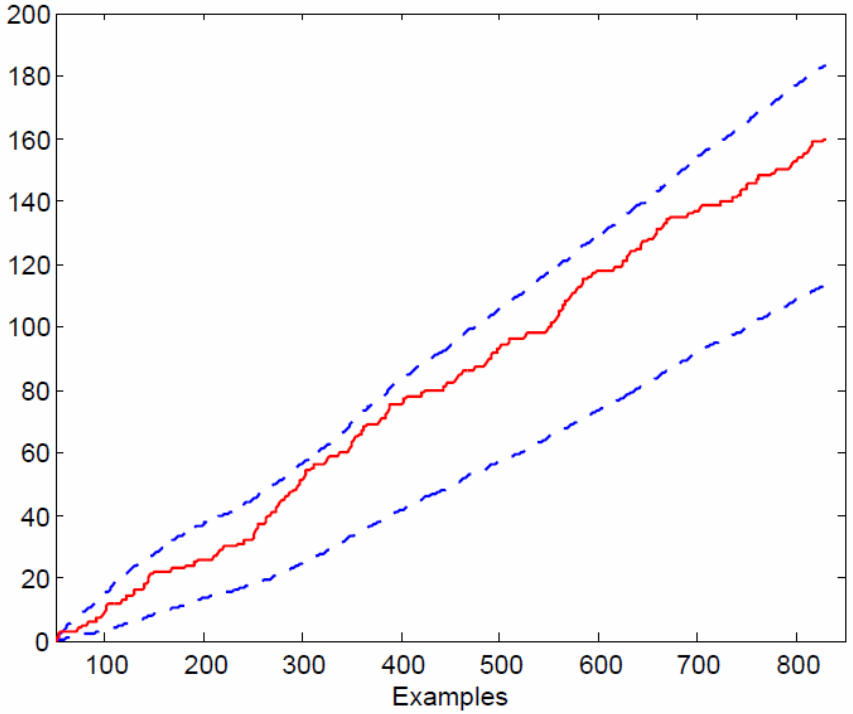
(a) Mammographic Mass



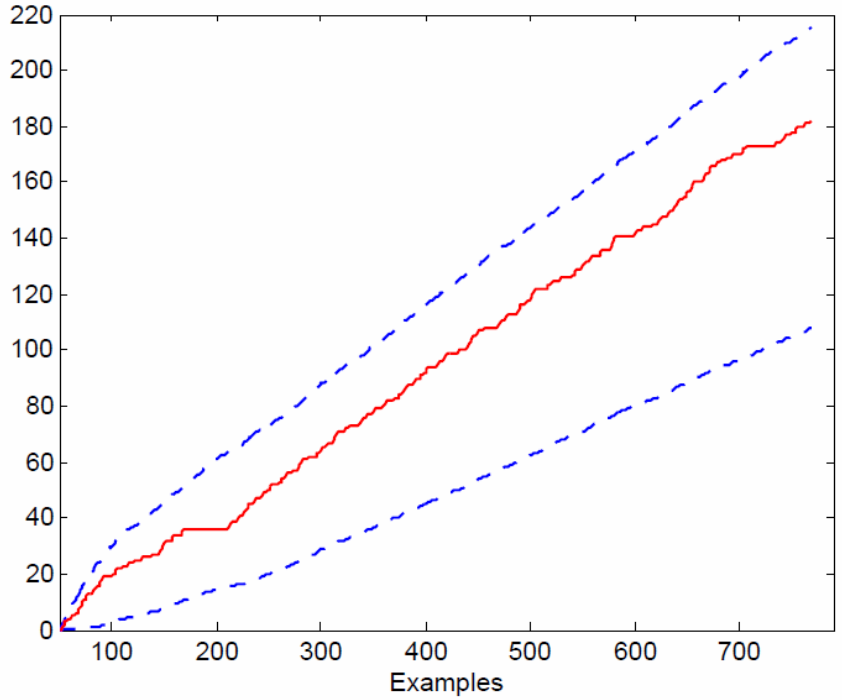
(b) Pima Indians Diabetes

On-line Experiments: NN-VP

- Both plots confirm that the probability intervals are well-calibrated



(a) Mammographic Mass

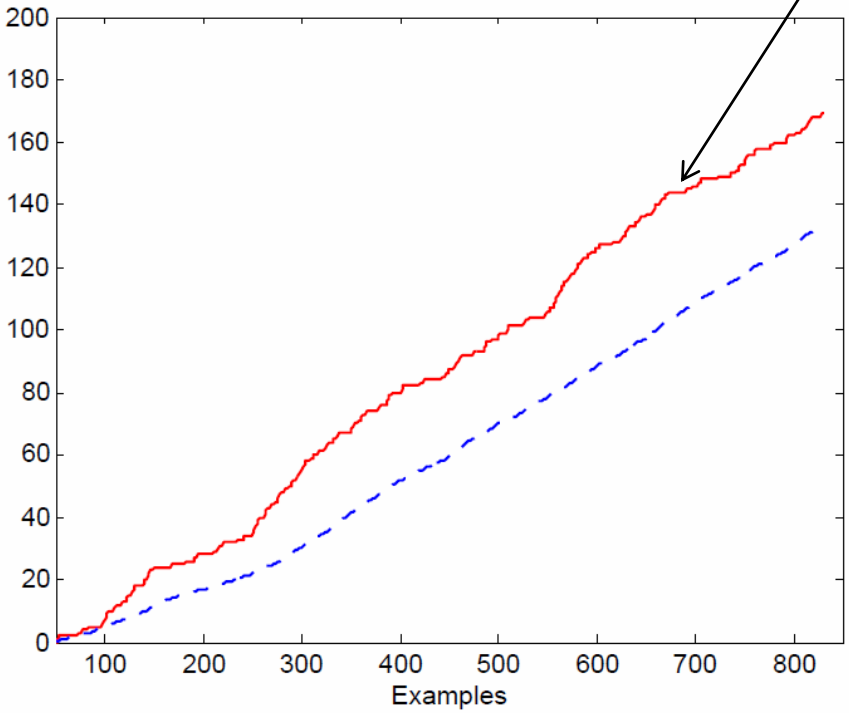


(b) Pima Indians Diabetes

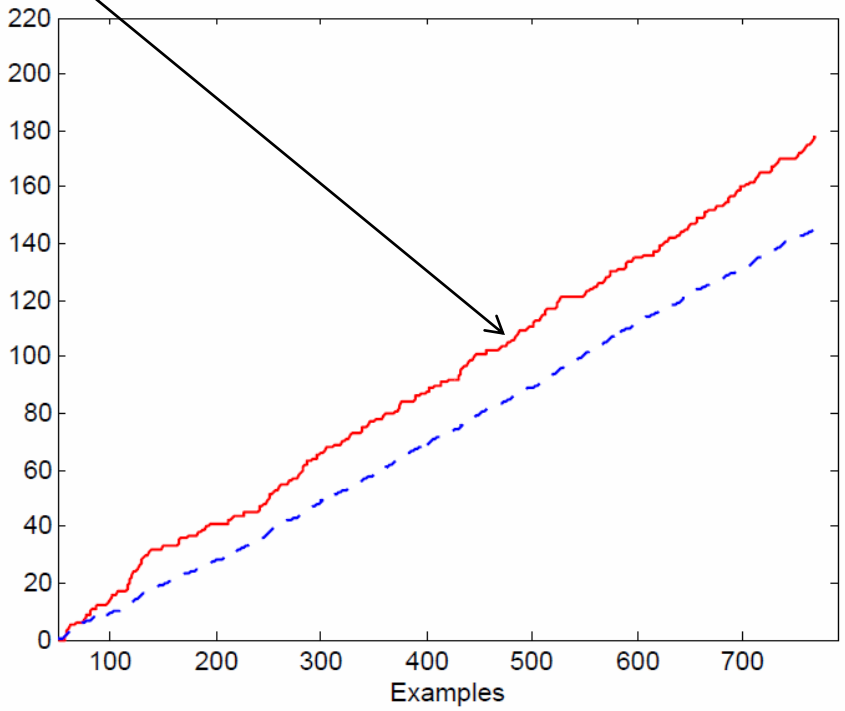
On-line Experiments: Conventional NN

Cumulative Error Curve

$$E_n = \sum_{i=1}^{n} err_i$$



(a) Mammographic Mass

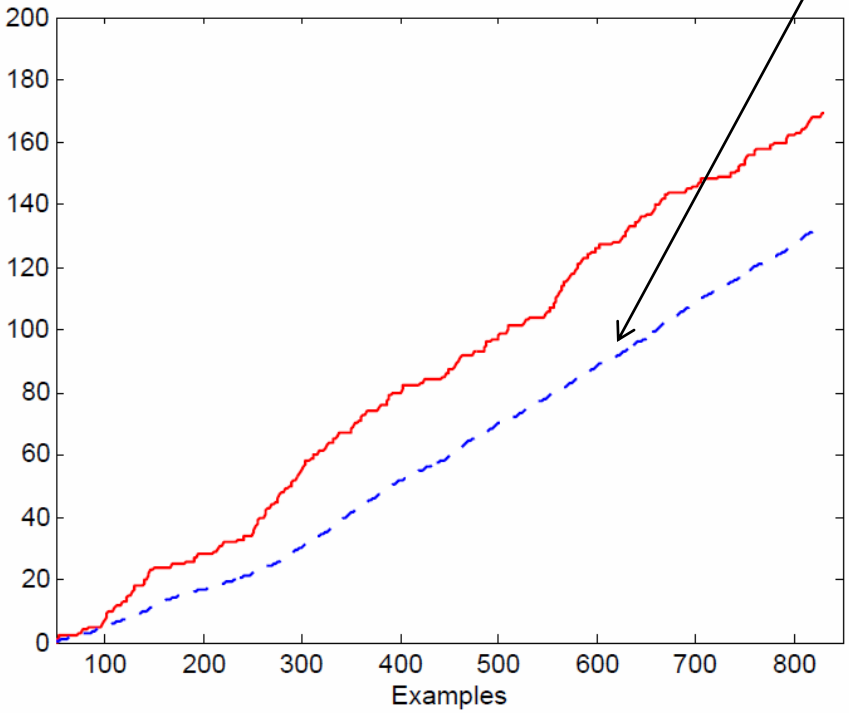


(b) Pima Indians Diabetes

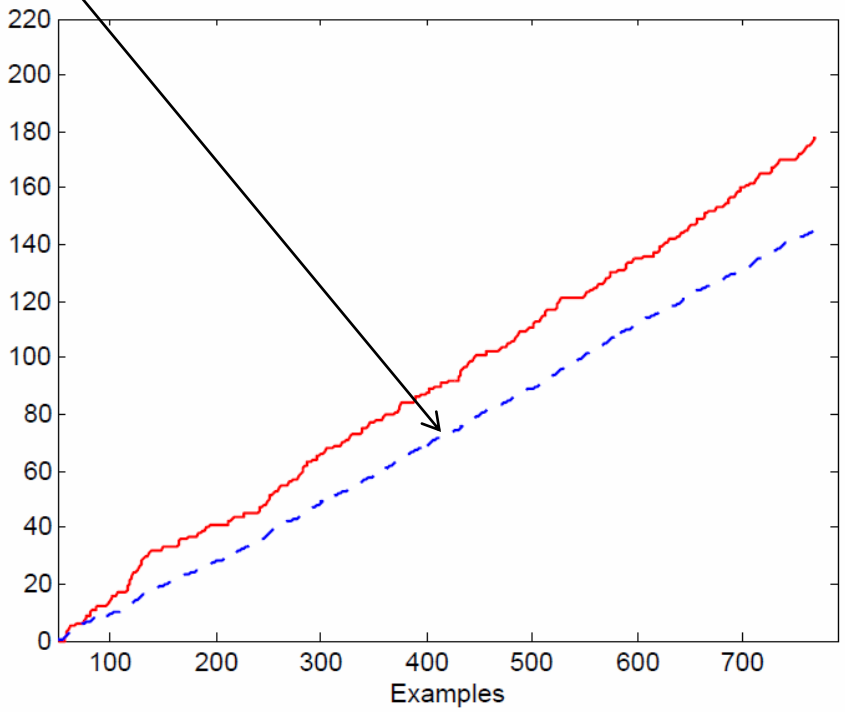
On-line Experiments: Conventional NN

Cumulative Error Probability Curve

$$EP_n = \sum_{i=1}^{n} |\hat{y}_i - \hat{p}_i|$$



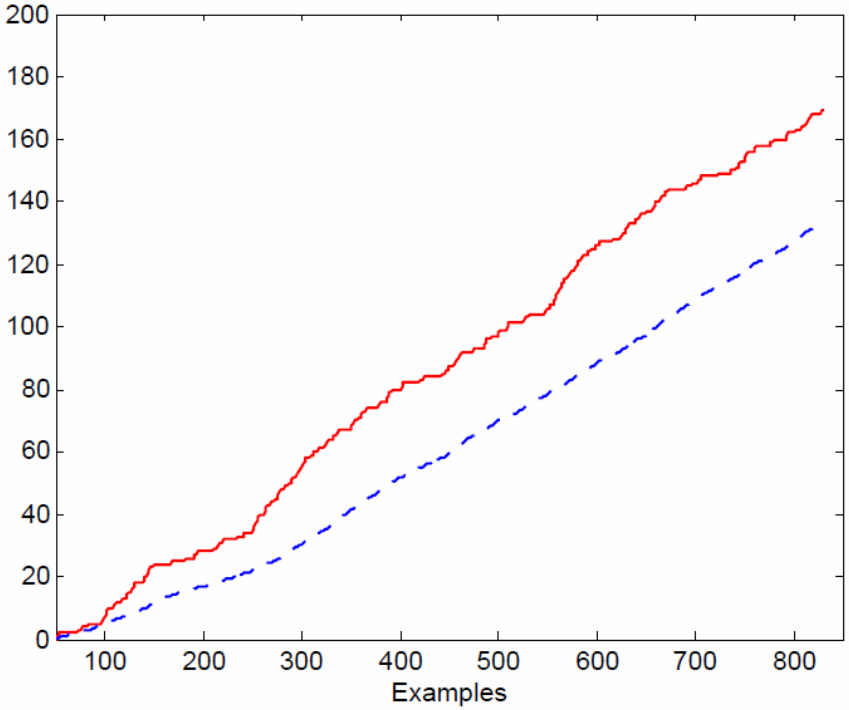
(a) Mammographic Mass



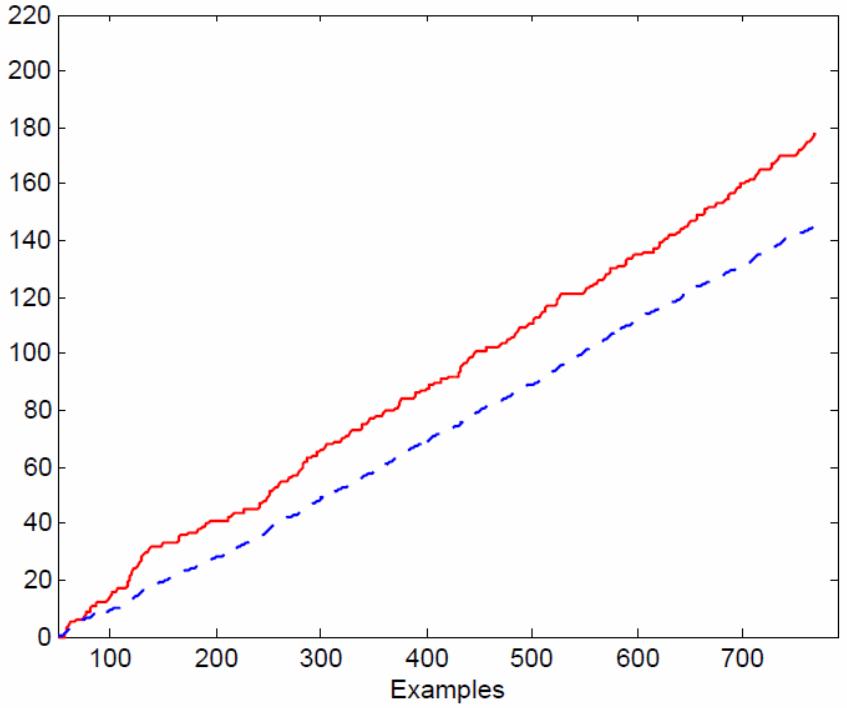
(b) Pima Indians Diabetes

On-line Experiments: Conventional NN

- The p-values of obtaining the resulting total number of errors given the probabilities produced by the NN were 0.000179 and 0.000548 respectively



(a) Mammographic Mass



(b) Pima Indians Diabetes

Batch Experiments

- Divided datasets into a training set of 200 examples and test set with all remaining examples
- 10 different random divisions

- The output of conventional NN is a single value:
  - the probability of the true class being 1
- For comparing the output of the NN-VP was converted to the corresponding single value:
  - the mean of $L(1)$ and $U(1)$

Batch Experiments: Quality Metrics

1. Accuracy
2. Cross-entropy error

$$CE = -\sum_{i=1}^{N} y_i \log(\hat{p}_i) + (1 - y_i) \log(1 - \hat{p}_i)$$

3. Brier Score

$$BS = \frac{1}{N} \sum_{i=1}^{N} (\hat{p}_i - y_i)^2$$

4. Reliability term of Brier Score

$$REL = \frac{1}{N} \sum_{k=1}^{K} n_k (r_k - \varphi_k)^2$$

Batch Experiments: Results

(a) Mammographic Mass

	Accuracy	CE	BS	REL
Conventional NN	78.83%	3298	0.1596	0.0040
NN-VP	78.92%	3054	0.1555	0.0023
Improvement (%)	0.11	7.40	2.57	42.50

(b) Pima Indians Diabetes

	Accuracy	CE	BS	REL
Conventional NN	74.56%	3084	0.1760	0.0074
NN-VP	74.26%	3014	0.1753	0.0035
Improvement (%)	-0.40	2.27	0.40	52.70

Conclusion

- Presented a Venn Predictor based on Neural Networks
- Unlike conventional NNs, NN-VP produces probability intervals, which are valid under the i.i.d. assumption
- The results obtained show that
  - the probability intervals of NN-VP are well-calibrated
  - the probabilities produced by conventional NNs can be misleading
  - even in terms of single probability values NN-VP is much more reliable than conventional NNs
- Future work
  - extension to multiclass NNs and evaluation of alternative Venn taxonomies
  - application of VP to other challenging problems

Acknowledgments

- The author is grateful to Professors Volodya Vovk and Alex Gammerman for useful discussions.

- This work was supported by the European Regional Development Fund and the Cyprus Government through the Cyprus Research Promotion Foundation "ΔΕΣΜΗ 2009-2010" research contract ΤΠΕ/OPIZO/0609(BIE)/24 ("Development of New Venn Prediction Methods for Osteoporosis Risk Assessment").

Thank you for listening