Application of Random Walks to Bayesian Classification and Business Decision Making

Clement Leung

School of Science and Engineering &

Guangdong Provincial Key Laboratory of Future Networks of Intelligence

The Chinese University of Hong Kong, Shenzhen, China

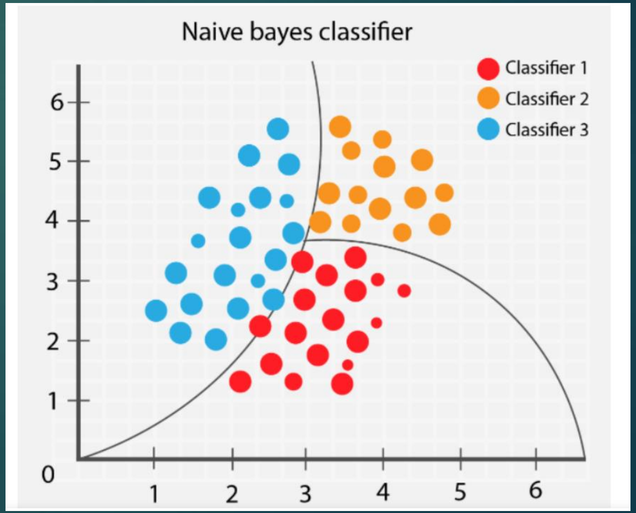
clementleung@cuhk.edu.cn

Clement LEUNG

- FULL PROFESSORSHIPS at
  - University of London, UK; National University of Singapore; Chinese University of Hong Kong, Shenzhen, China; Hong Kong Baptist University; Victoria University, Australia
- Two US patents, five books and over 150 research articles
- Program Chair, Keynote Speaker, Panel Expert of major International Conferences
- Editorial Board of ten International Journals
- Listed in Who's Who in the World and Great Minds of the 21st Century
- Fellow of the British Computer Society, Fellow of the Royal Society of Arts, and Fellow of the International Academy, Research, and Industry Association

Classification Problems are Ubiquitous

- Many classifiers are applied to the same object
- Many objects are being classified



Classification Problems are Pervasive in Business

- Should we adopt this advertising channel or not?
- Should we include this particular product in our promotion this month?
- Should we offer employment to this applicant?

Employee Performance Appraisal: multiple assessors of multiple employees

	Manager 1	Manager 2	Manager 3
Employee 1	Acceptable	Not Acceptable	Acceptable
Employee 2	Acceptable	……	
……			
Employee N	Not Acceptable	Acceptable	……

Medical Treatment: multiple physicians assessing multiple patients

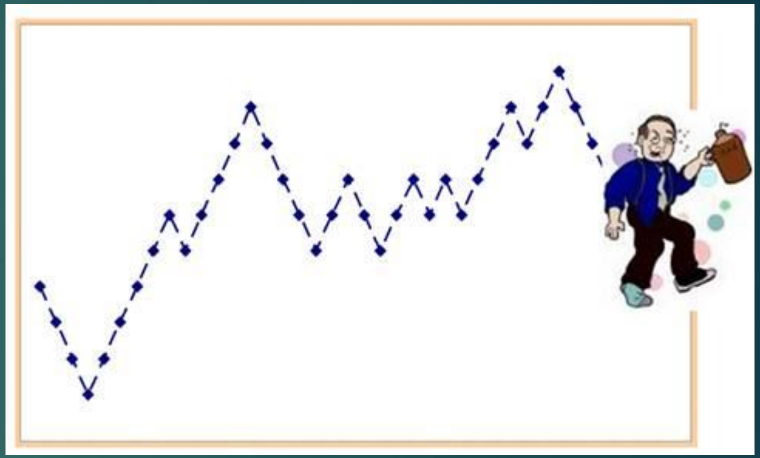
	Physician 1	Physician 2	Physician 3
Patient 1	Invasive operation	No surgery	No surgery
Patient 2	Invasive operation	……	
……			
Patient N	No Surgery	No surgery	……

One-Dimension Random Walk

- Task *i*
  - corresponds to object *i*
- Predictor *j*
  - corresponds to classifier *j*
- A set of classification labels Z_{ij}, where

$$Z_{ij} = \begin{cases} -1 \\ \\ +1 \end{cases}$$

  is a binary label taking on the values +1 or -1.
- A +1 classification label can be regarded as taking a step to the right, while a -1 label can be regarded as taking a step to the left



Displacement of the Random Walk

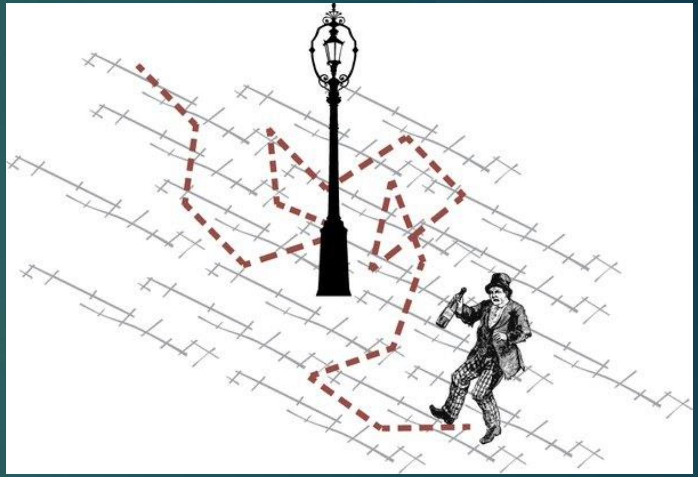
From the set of independent identically distributed random variables $\{Z_{ij}\}_{j>0}$ with

$$\mathbb{P}[Z_{ij} = +1] = p_j$$
$$\mathbb{P}[Z_{ij} = -1] = q_j$$

where $p_j + q_j = 1$, the displacement of the random walk after n steps, which corresponds to the outcome of n cumulative classification results, for a given task i is given by

$$X_{in} = \sum_{j=1}^{n} Z_{ij}$$

where it is assumed that $X_{i0} = 0$.



Ground Truth

- For a total of *M* tasks (*M* random walks), we want to determine the error of the ground truth vector of the problem

$$\boldsymbol{g} = \begin{pmatrix} g_1 \\ g_2 \\ \cdots \\ \cdots \\ g_M \end{pmatrix}$$

where the elements g_i can take on the value +1 or -1

Naïve Bayes

- We adopt the Naïve Bayes property that the predictors are independent

Predicted Class

- For task i, we assume that a fixed number of classifiers n_i are used to complete the classification task, after which majority voting determines the class
  - n_i is normally assumed to be odd to avoid an equal number of votes for each class being received
- n_i steps are taken
  - n_i can be regarded as a constraint placed on the budget
  - the total budget for the M tasks is $n_1 + n_2 + \ldots + n_M$
- Denote by $\hat{g}_i$ the predicted class for task i, and by

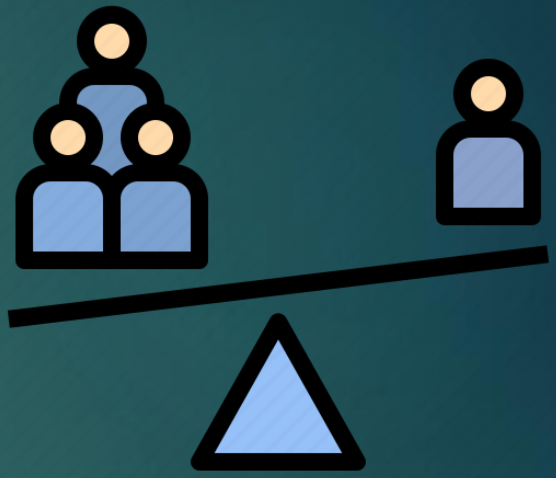
$$\hat{\boldsymbol{g}} = \begin{pmatrix} \hat{g}_1 \\ \hat{g}_2 \\ \ldots \\ \ldots \\ \hat{g}_M \end{pmatrix}$$

the predicted class vector

Majority Vote as Random Walk Displacement

For any given task *i*,

(i) the ground truth for the task is −1 when $q_i > p_i$, and
(ii) the ground truth for the task is +1 when $p_i > q_i$.



<u>Proof:</u>

Taking expectations of the net displacement

$$E(X_{in}) = \sum_{j=1}^{n} E\left(Z_{ij}\right) = \sum_{j=1}^{n} (p_i - q_i) = n(p_i - q_i)$$

As $n \to \infty$, when $q_i > p_i$, the mean displacement will drift to $-\infty$, indicating the majority of the votes are for the class −1, which completes the proof of (i). Similar argument applies to the case $q_i < p_i$, resulting in the majority of the votes are for the class +1.

Displacement Properties

Denoting X_{in_i} by X_{n_i}, since X_{n_i} is sufficient to indicate that the task in question is task *i*, it can be shown that

(i) For *k* an even integer,

$$\mathbb{P}\left[X_{n_i} = k\right] = 0.$$

(ii) For *k* an odd integer,

$$\mathbb{P}\left[X_{n_i} = k\right] = \binom{n_i}{\frac{n_i + k}{2}} p_i^{\frac{n_i+k}{2}} q_i^{\frac{n_i-k}{2}}$$

Prediction Error

- A prediction error will result if $p_i > q_i$, yet the final position of the walk lands in the negative axis
  - In the long run, if $p_i > q_i$, the net drift will be to the right and so the ground truth should be +1
- A prediction error will result if $q_i > p_i$, yet the final position of the walk lands in the positive axis
  - In the long run, if $q_i > p_i$, the net drift will be to the left and so the ground truth should be -1

Probability of Prediction Error

By analyzing the random walk behaviour, the error probability can be shown to be

$$\mathbb{P}[\hat{g} \neq g] = 1 - \prod_{j \,\epsilon\, P} \{1 - \sum_{k \,\epsilon\, \Omega^-} \binom{n_j}{\frac{n_j + |k|}{2}} p_j^{\frac{n_j - |k|}{2}} q_j^{\frac{n_j + |k|}{2}} \} \prod_{i \,\epsilon\, Q} \{1 - \sum_{k \,\epsilon\, \Omega^+} \binom{n_i}{\frac{n_i + k}{2}} p_i^{\frac{n_i + k}{2}} q_i^{\frac{n_i - k}{2}} \} .$$

where,

P is the set of indexes of tasks with ground truth equalled to +1,
Q is the set of indexes of tasks with ground truth equalled to −1,
$\Omega^+ = \{2n - 1\}_{n=1}^{\frac{n_i - 1}{2}}$ is the set of positive odd integers from 1 to n_i (inclusive of 1 and n_i),
$\Omega^- = \{1 - 2n\}_{n=1}^{\frac{n_j - 1}{2}}$ is the set of negative odd integers from -1 to -n_j (inclusive of -1 and -n_j).

Exact and Approximate Error Bounds for Ground Truth Class -1

For any task *i* with a ground truth class of −1, we have

$$\mathbb{P}[\hat{g}_i \neq g_i] \leq \left\lceil \frac{n_i}{2} \right\rceil \binom{n_i}{\frac{n_i + \lfloor (n_i+1)p_i \rfloor}{2}} p_i^{\frac{n_i+\lfloor (n_i+1)p_i \rfloor}{2}} q_i^{\frac{n_i-\lfloor (n_i+1)p_i \rfloor}{2}}$$

and to simplify the above calculations, we can use the approximation

$$\mathbb{P}[\hat{g}_i \neq g_i] \lesssim \frac{\Gamma(n_i + 2)}{2\Gamma(\frac{n_i(1+p_i)}{2} + 1)\Gamma(\frac{n_i q_i}{2} + 1)} p_i^{\frac{n_i(1+p_i)}{2}} q_i^{\frac{n_i q_i}{2}}$$

where Γ(.) is the gamma function.

Exact and Approximate Bounds for Ground Truth Class +1

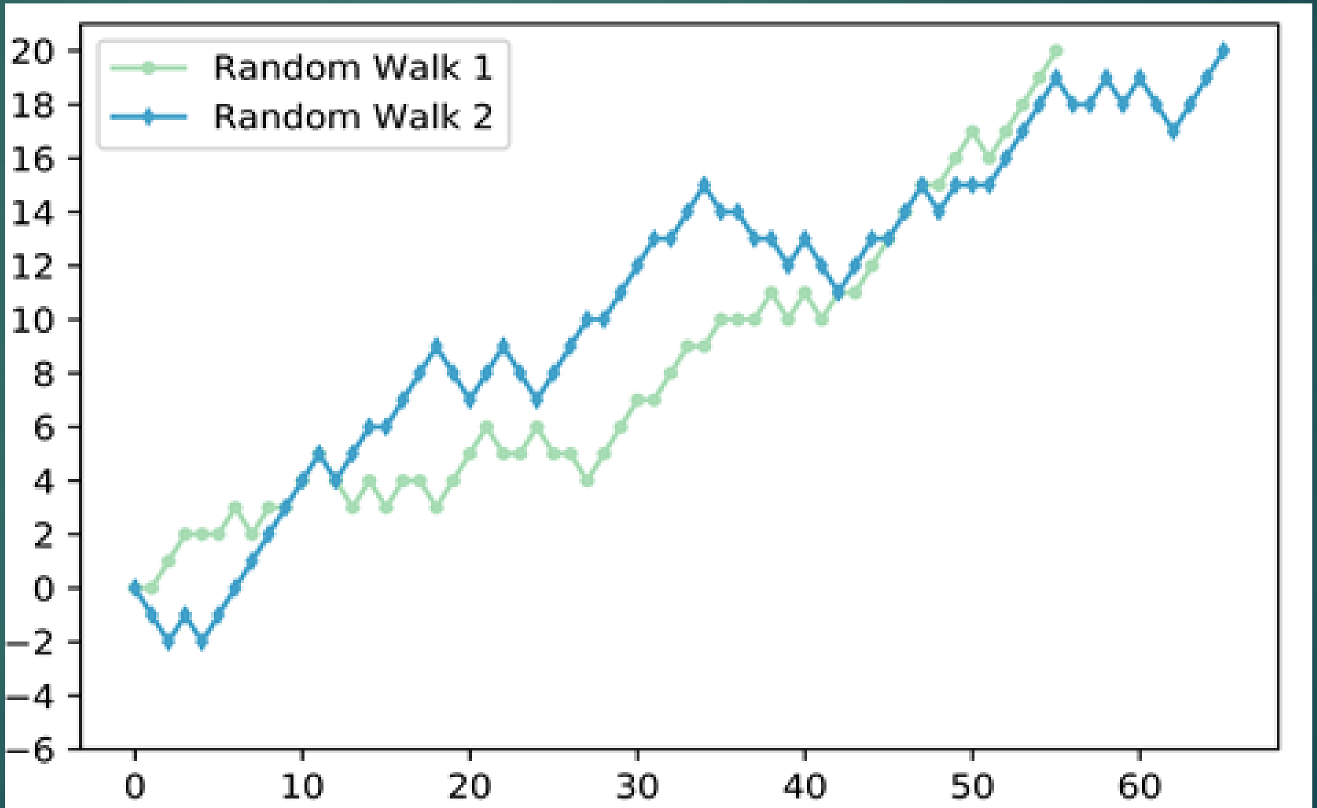
For any task *j* with a ground truth class of +1, we have

$$\mathbb{P}[\hat{g}_j \neq g_j] \leq \left\lceil \frac{n_j}{2} \right\rceil \binom{n_j}{\frac{n_j + \lfloor (n_j+1) q_j \rfloor}{2}} p_j^{\frac{n_j - \lfloor (n_j+1) q_j \rfloor}{2}} q_j^{\frac{n_j + \lfloor (n_j+1) q_j \rfloor}{2}}$$

and the corresponding approximation is

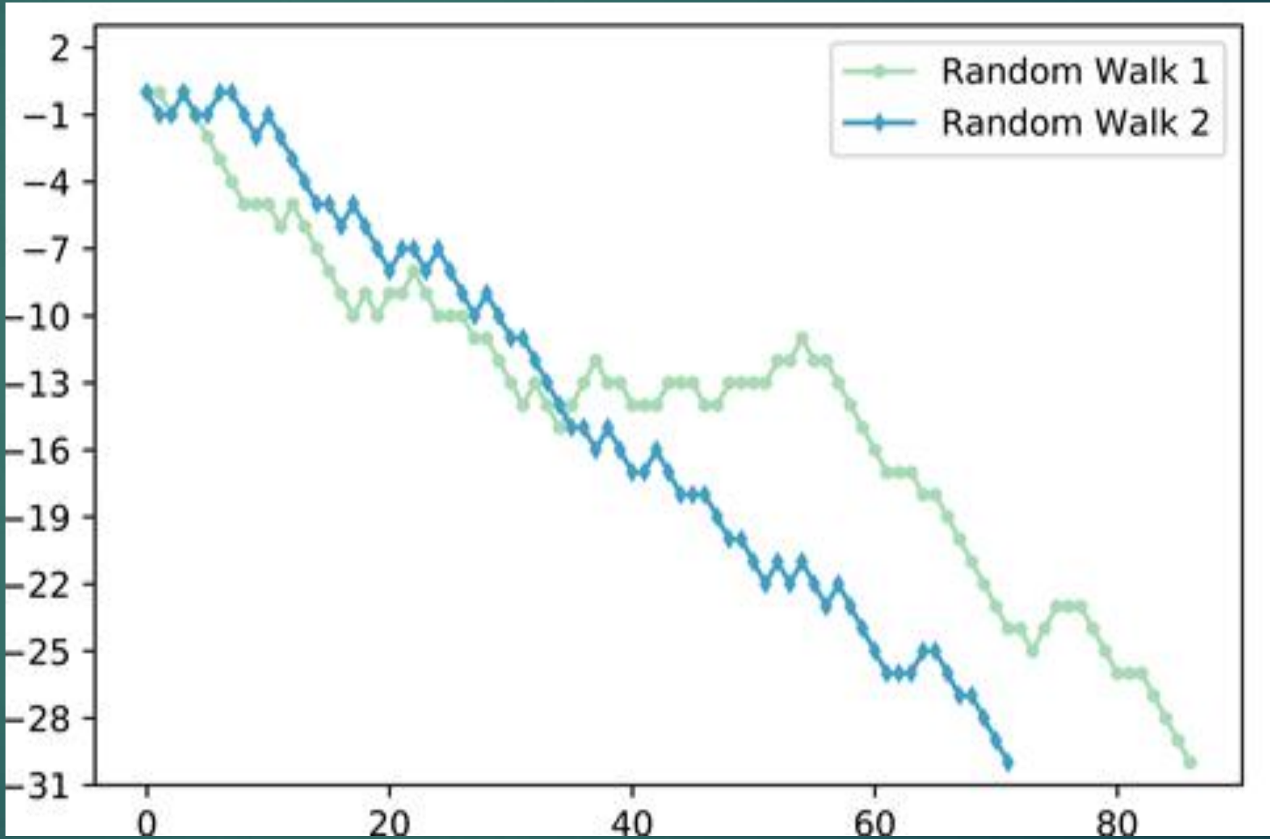
$$\mathbb{P}[\hat{g}_j \neq g_j] \lesssim \frac{\Gamma(n_j + 2)}{2\Gamma(\frac{n_j(1 + q_j)}{2} + 1)\Gamma(\frac{n_j p_j}{2} + 1)} p_j^{\frac{n_j p_j}{2}} q_j^{\frac{n_j(1+q_j)}{2}}$$

Random Walk Simulation Experiments



Random Walks with Net Positive Drift

Random Walk Simulation Experiments



Random Walks with Net Negative Drift

Comparison of Theoretical and Experimental Results

Each set of parameter settings are run 100,000 times.

Observed absolute errors are < 2%

q	*p*	No. of classif-iers *n*	No. of times landing on +ve axis	Obs Freq of Error	Th Freq of Error	% Error Between Th & Obs
0.6	0.4	1	40100	0.401	0.400	0.25
0.6	0.4	3	35159	0.35159	0.352	-0.12
0.6	0.4	5	31720	0.3172	0.317	-0.08
0.6	0.4	7	28753	0.28753	0.290	-0.79
0.6	0.4	9	26883	0.26883	0.267	0.84
0.6	0.4	11	24391	0.24391	0.247	-1.06
0.6	0.4	13	22896	0.22896	0.229	0.05
0.6	0.4	15	21138	0.21138	0.213	-0.80
0.7	0.3	1	29874	0.29874	0.300	-0.42
0.7	0.3	3	21481	0.21481	0.216	-0.55
0.7	0.3	5	16362	0.16362	0.163	0.33
0.7	0.3	7	12620	0.1262	0.126	0.13
0.7	0.3	9	9886	0.09886	0.099	0.05
0.7	0.3	11	7909	0.07909	0.078	1.09
0.7	0.3	13	6328	0.06328	0.062	1.43

Observed Classification Errors and Comparison with Theoretical Results

Summary and Conclusion

- Multiple classification problems are ubiquitous in business decision making
- Classification errors are unavoidable and cannot always be eliminated
  - the occurrences of false positives and false negatives are common due to limited accuracies in the underlying classifiers
- In many practical situations, it is unrealistic to assume that absolute and objective ground truth classes are available
  - the multiple classification problem is studied using the Naïve Bayes approach, where the ground truth is not absolute and is determined by the view of the majority of classifiers.

Summary and Conclusion

- The penalty of misclassification is substantial and cannot be disregarded
  - Ideally, all classifiers should applied to obtain a classification decision, but resource and time constraints often make this impractical, and classification decisions will have to be made within finite time points prior to fully exhaustive classification
- We make use of a random walk model to study the situation and have derived closed-form expressions for the probability of error as well as useful error bounds as a function of the budget constraint.

Summary and Conclusion

- We find that by raising the budget, the probability of error in classification can be lowered
  - the extent of the improvement can be suitably quantified and controlled
- Extensive experiments have been performed
  - the results of which show good agreement with the theoretical results

Thank you!