Sleeping Experts in On-line Prediction

Yuri Kalnishkan

Department of Computer Science
and Computer Learning Research Centre
Royal Holloway, University of London

November 2014

Contents

Protocol

- we try to predict elements of a sequence $\omega_1, \omega_2, \omega_3, \ldots \in \Omega$
- we output predictions $\gamma_1, \gamma_2, \gamma_3, \ldots \in \Gamma$
- we can make use of signals $s_1, s_2, s_3 \ldots$
- protocol:

  FOR $t = 1, 2, \ldots$
  (1) Learner observes signal s_t
  (2) Learner chooses a prediction $\gamma_t \in \Gamma$
  (3) Learner observes the actual outcome $\omega_t \in \Omega$
  (4) Learner suffers loss $\lambda(\omega_t, \gamma_t)$
  END FOR
- loss over T trials sums up to the cumulative loss

$$\text{Loss}(\omega_1, \omega_2, \ldots, \omega_T) = \sum_{i=1}^{T} \lambda(\omega_i, \gamma_i)$$

Formalisation

- a *game* $\mathfrak{G}$ is a triple $\langle \Omega, \Gamma, \lambda \rangle$
  - ◇ Ω is the *outcome space*
  - ◇ Γ is the *prediction space*
  - ◇ $\lambda : \Omega \times \Gamma \to [0, +\infty]$ is the *loss function*
- important special case: binary games
  - ◇ $\Omega = \mathbb{B} = \{0, 1\}$
  - ◇ $\Gamma = [0, 1]$

Examples

- square-loss game: $\Omega = \{0, 1\}$, $\Gamma = [0, 1]$, $\lambda(\omega, \gamma) = (\omega - \gamma)^2$
- absolute-loss game: $\Omega = \{0, 1\}$, $\Gamma = [0, 1]$, $\lambda(\omega, \gamma) = |\omega - \gamma|$
- logarithmic game: $\Omega = \{0, 1\}$, $\Gamma = [0, 1]$

$$\lambda(\omega, \gamma) = \begin{cases} -\log_2(1-\gamma) & \text{if} \quad \omega = 0 \\ -\log_2 \gamma & \text{if} \quad \omega = 1 \end{cases}$$

— can take the value $+\infty$
- simple prediction game: $\Omega = \Gamma = \{0, 1\}$

$$\lambda(\omega, \gamma) = \begin{cases} 0 & \text{if} \quad \omega = \gamma \\ 1 & \text{if} \quad \omega \neq \gamma \end{cases}$$

Experts

- suppose that we can see predictions of N Experts $E_1, E_2, \ldots, E_N$

  FOR $t = 1, 2, \ldots$
  (1) Experts output $\gamma_t^n \in \Gamma$, $n = 1, \ldots, N$
  (2) Learner outputs $\gamma_t \in \Gamma$
  (3) the outcome $\omega_t \in \Omega$ occurs
  (4) Learner suffers loss $\lambda(\gamma_t, \omega_t)$
  (5) Experts suffer losses $\lambda(\gamma_t^n, \omega_t)$, $n = 1, 2, \ldots, N$
  END FOR
- we want to be sure not to suffer loss much greater than that of the best expert
  — i.e., we want a guarantee of the type $\mathrm{Loss}(T) \lesssim \mathrm{Loss}_{E_n}(T)$ for all n and T

Aggregating Algorithm

- takes a parameter $\eta > 0$ (learning rate)
- maintains weights w_t^n for experts
  — they are initialised with a distribution $q_1, q_2, \ldots, q_N$ (e.g., uniform $q_n = 1/N$)
  — after expert E_n suffers loss $\lambda(\gamma_t^n, \omega_t)$, its loss is updated as

$$w_{t+1}^n = w_t^n e^{-\eta \lambda(\gamma_t^n, \omega_t)}$$

- on step t normalised weights $p_t^n = w_t^n / \sum_{m=1}^N w_t^m$ are used to work out Learner's prediction γ_t satisfying

$$\lambda(\gamma_t, \omega) \leq -C(\eta) \frac{1}{\eta} \ln \sum_{n=1}^N p_t^n e^{-\eta \lambda(\gamma_t^n, \omega)}$$

  for all $\omega \in \Omega$
- $C(\eta)$ is the smallest number such that γ_t can always be found

Algorithm

parameters: η and initial distribution $q_1, q_2, \ldots, q_N$

(1) initialise $w_1^n = q_n$, $n = 1, 2, \ldots, N$
FOR $t = 1, 2, \ldots$
  (2) read experts' predictions γ_t^n, $n = 1, 2, \ldots, N$
  (3) normalise weights $p_t^n = w_t^n / \sum_{n=1}^N w_t^n$
  (4) solve the system ($\omega \in \Omega$):
    $\lambda(\gamma, \omega) \le -\frac{C(\eta)}{\eta} \ln \sum_{n=1}^N p_t^n e^{-\eta\lambda(\gamma_t^n, \omega)}$
    w.r.t. γ and output γ_t
  (5) observe ω_t
  (6) update experts' weights $w_{t+1}^n = w_t^n e^{-\eta\lambda(\gamma_t^n, \omega)}$, $n = 1, 2, \ldots, N$
END FOR

The Guarantees

- for all sequences of outcomes $\omega_1, \omega_2, \ldots$ and experts' predictions and every $n = 1, 2, \ldots, N$ we get

$$\mathrm{Loss}(t) \le C(\eta)\,\mathrm{Loss}_{E_n}(t) + \frac{C(\eta)}{\eta} \ln \frac{1}{q_n}$$

  — if the initial weights are uniform

$$\mathrm{Loss}(t) \le C(\eta)\,\mathrm{Loss}_{E_n}(t) + \frac{C(\eta)}{\eta} \ln N$$

- if $C(\eta) =$ for some η, the game is called *mixable*
  — for mixable games we get

$$\mathrm{Loss}(t) \le \mathrm{Loss}_{E_n}(t) + \frac{\ln N}{\eta}$$

- the coefficients $C(\eta)$ and $C(\eta)/\eta$ are optimal

Definition

- suppose that an expert can skip turns...
- a *specialist expert* can refrain from prediction on step t
  — if it makes a prediction, we say that it is *awake*
  — otherwise we say that it *sleeps*
  — specialist experts *can sleep* a sleeping expert *is sleeping now*
- how can we handle them?
  — idea: let a sleeping expert follow awake ones
- literature:
  — originated in [Y. Freund et al, Using and combining predictors that specialize, Proceedings of STOC 1997]
  — we follow [A.Chernov and V.Vovk, Prediction with expert evaluators' advice, Proceedings of ALT 2009]

Algorithm (1)

- look at

$$e^{-\eta\lambda(\gamma_t,\omega)} \geq \sum_{n=1}^{N} p_t^n e^{-\eta\lambda(\gamma_t^n,\omega)}$$

— this is the exponentiated key inequality of the aggregating algorithm for the mixable case

- let us single out the terms corresponding to sleeping and awake experts

$$e^{-\eta\lambda(\gamma_t,\omega)} \geq \sum_{n: E_n \text{ is awake}} p_t^n e^{-\eta\lambda(\gamma_t^n,\omega)} + \sum_{n: E_n \text{ sleeps}} p_t^n e^{-\eta\lambda(\gamma_t,\omega)}$$

Algorithm (2)

- the sum over sleeping experts cancels out

$$e^{-\eta\lambda(\gamma_t,\omega)} \geq \frac{1}{Z_t} \sum_{n: E_n \text{ is awake}} p_t^n e^{-\eta\lambda(\gamma_t^n,\omega)}$$

where Z_t is the weight of experts awake on step t

- specialist experts can be handled with a minimum modification of the algorithm:
— the summation is done over awake experts and their weights are normalised to 1
— the sleeping experts output "the crowd's" prediction γ_t, and so their weights are updated as $w_{t+1}^n = w_t^n e^{-\eta\lambda(\gamma_t,\omega_t)}$ using the Learner's γ_t

Loss Bound

- in

$$\mathrm{Loss}(T) \leq \mathrm{Loss}_{E_n}(T) + \frac{1}{\eta}\ln N$$

we can drop the terms where E_n sleeps

- we get

$$\overline{\mathrm{Loss}}^n(T) \leq \overline{\mathrm{Loss}}^n_{E_n}(T) + \frac{1}{\eta}\ln N$$

where the sum in $\overline{\mathrm{Loss}}^n$ is taken over steps where E_n was awake

Discussion

- new experts can be added on-the-fly
— e.g., a new expert can start predicting upon completing a training stage
- the weight of a new expert joining at time T can be worked out using our loss $\mathrm{Loss}(T-1)$, because it was following us while it was sleeping
- sleeping experts can be used to extract relevant historical information

Options

- an option on a share is a contract of the following kind:

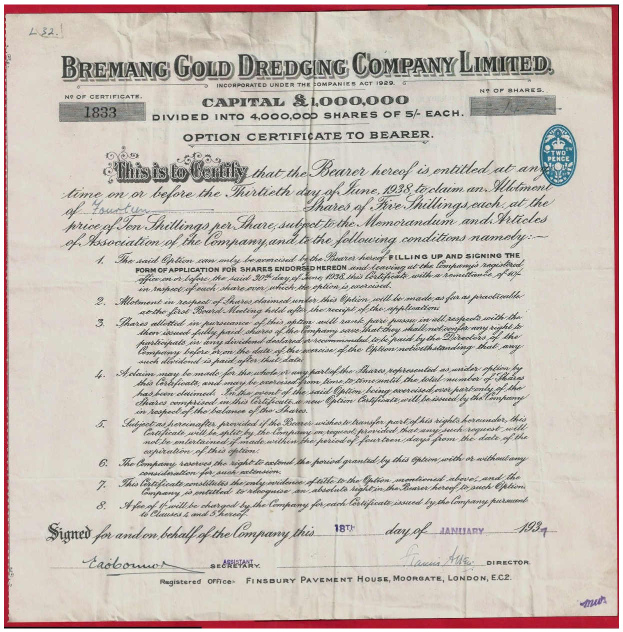
The bearer of this may buy a share of ABC ltd in December 2014 at a price of $10.

buy: an option that entitles its holder to buy at a fixed price is called *call*, and an option that entitles to sell is called *put*

share: the financial instrument that is bought or sold is called *the underlying (asset)*; it can be a share, a futures, an index etc.

Dec 2014: the option shows the date when it can be *exercised*, i.e., when the holder may use it; it is called *maturity* or *expiration* (it makes sense to exercise this option if and only if the share price in December 2014 exceeds $10)
— the stock exchange usually fixes four expiration dates in a year and only allows options with those expiration dates

$10: the fixed price written in the option is called *strike*



L 32.

BREMANG GOLD DREDGING COMPANY LIMITED.

INCORPORATED UNDER THE COMPANIES ACT 1929.

Nº OF CERTIFICATE. 1833

CAPITAL £1,000,000

DIVIDED INTO 4,000,000 SHARES OF 5/- EACH.

Nº OF SHARES. 14

OPTION CERTIFICATE TO BEARER.

TWO PENCE

This is to Certify that the Bearer hereof is entitled at any time on or before the Thirtieth day of June, 1938 to claim an Allotment of Fourteen Shares of Five Shillings each, at the price of Ten Shillings per Share, subject to the Memorandum and Articles of Association of the Company, and to the following conditions namely:—

1. The said Option can only be exercised by the Bearer hereof FILLING UP AND SIGNING THE FORM OF APPLICATION FOR SHARES ENDORSED HEREON and leaving at the Company's registered office on or before the said 30th day of June, 1938, this Certificate with a remittance of 10/- in respect of each share over which the option is exercised.
2. Allotment in respect of Shares claimed under this Option will be made, as far as practicable at the first Board Meeting held after the receipt of the application.
3. Shares allotted in pursuance of this option will rank pari passu in all respects with the then issued fully paid shares of the Company save that they shall not confer any right to participate in any dividend declared or recommended to be paid by the Directors of the Company before or on the date of the exercise of the Option notwithstanding that any such dividend is paid after that date.
4. A claim may be made for the whole or any part of the Shares represented as under option by this Certificate, and may be exercised from time to time until the total number of Shares has been claimed. In the event of the said Option being exercised over part only of the Shares comprised in this Certificate, a new Option Certificate will be issued by the Company in respect of the balance of the Shares.
5. Subject as hereinafter provided if the Bearer wishes to transfer part of his rights hereunder, this Certificate will be split by the Company on request, provided that any such request will not be entertained if made within the period of fourteen days from the date of the expiration of this option.
6. The Company reserves the right to extend the period granted by this Option with or without any consideration for such extension.
7. This Certificate constitutes the only evidence of title to the Option mentioned above, and the Company is entitled to recognise an absolute right in the Bearer hereof to such Option.
8. A fee of 1/- will be charged by the Company for each Certificate issued by the Company pursuant to Clauses 4 and 5 hereof.

Signed for and on behalf of the Company this 18TH day of JANUARY 1937

ASSISTANT SECRETARY. DIRECTOR.

Registered Office:- FINSBURY PAVEMENT HOUSE, MOORGATE, LONDON, E.C.2.

Mathematical Interpretation

- call option:

The bearer of this is entitled to the sum of $\max(S_T - X, 0)$ at the moment T.

- put option:

The bearer of this is entitled to the sum of $\max(X - S_T, 0)$ at the moment T.

- here:
  T – expiration moment
  S_T – the price of the underlying at time T
  X – strike price

Black-Scholes Formulas

- what is the value of an option?
- call: $c = S\Phi(d_1) - Xe^{-rT}\Phi(d_2)$
- put: $p = Xe^{-rT}\Phi(-d_2) - S\Phi(-d_1)$

$$d_1 = (\ln(S/X) + (r + \sigma^2/2)T)/(\sigma\sqrt{T})$$
$$d_2 = (\ln(S/X) + (r - \sigma^2/2)T)/(\sigma\sqrt{T})$$

where:
X – strike
T – time of expiration/maturity
S – the price of the underlying
r – interest rate (often taken to be 0)
σ – volatility
Φ – the distribution function of the Gaussian distribution

Volatility

- volatility is the only parameter that is not observed directly
- the Black-Scholes(-Merton) theory assumes that the logarithm of the stock price $\ln S_t$ follows the generalised Brownian motion so that the variance of $(\ln S_{t+\Delta t} - \ln S_t)$ is $\sigma^2 \Delta t$
- the volatility σ can be estimated statistically from share price observations

Implied Volatility (1)

- the option price can actually be observed (we can see the quote)
- let us use the B-S formula the other way round and work out the volatility from the option price
- this estimate is called *implied volatility*

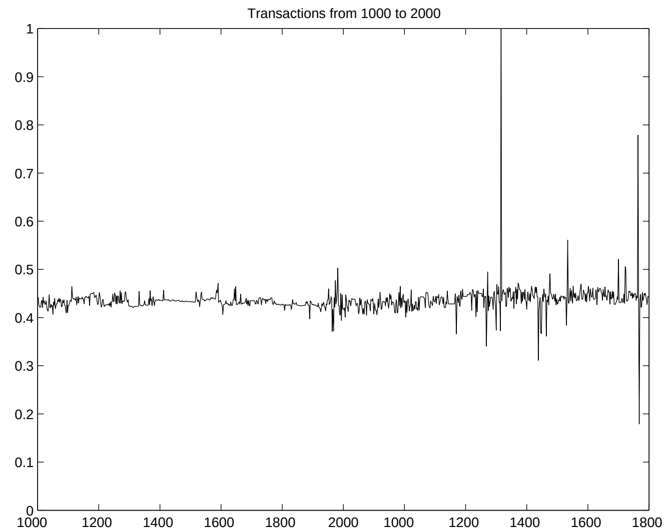
Implied Volatility (2)

- the Black-Scholes volatility is specified by the share and therefore does not depend on the option parameters
- in practice the implied volatility does and we get a function $\sigma(X, T)$
  — there is no unique commonly recognised explanation to this
- the graph of $\sigma(X)$ for fixed T is called *volatility smile*
- implied volatility is a commonly used and intuitive (for traders) parameter

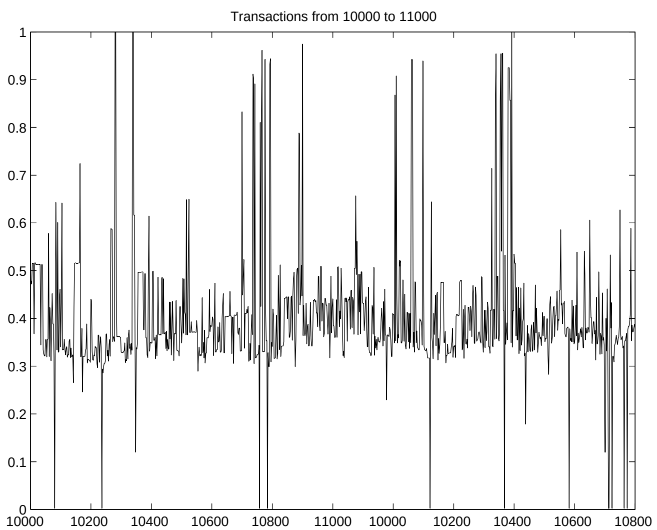
Predicting Implied Volatility

- we are given a log of option transactions (on a fixed underlying with a fixed maturity)
— the data was provided by the Russian Trading System Stock Exchange
- we want to predict implied volatility for the next transaction
— we can use the current stock price, strike, and time to maturity (the interest rate is assumed to be 0) but not the option price (as it immediately implies the volatility)
- we use square loss
— this is a mixable game (though not binary)
- consider shares of Russian Energy Systems maturing in December 2006 (about 13000 transactions)

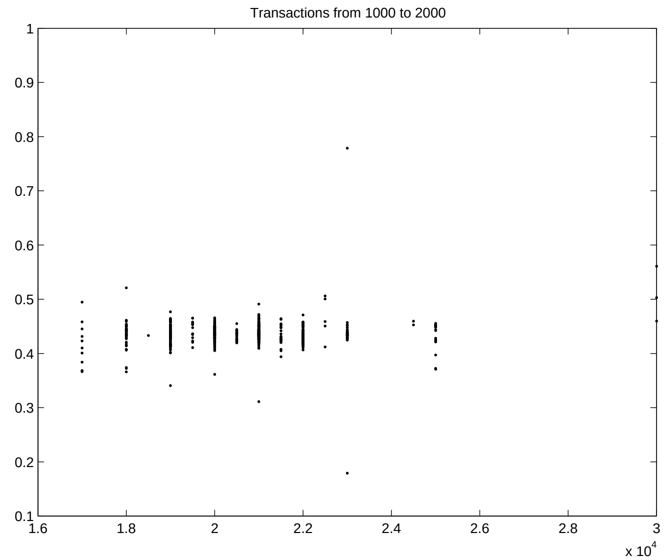
Volatility vs Transaction Number, 1000-2000



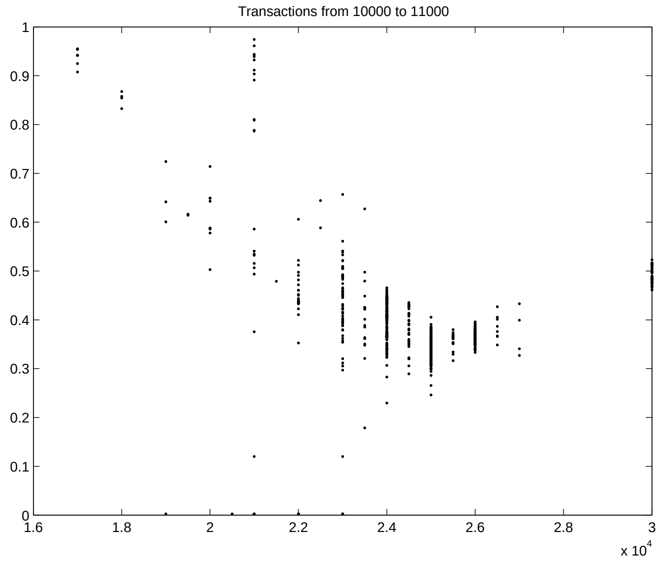
Volatility vs Transaction Number, 10000-11000



Volatility vs Strike, 1000-2000



Volatility vs Strike, 10000-11000



Naive Algorithm

- let us predict implied volatility as the implied volatility from the previous transaction with the same strike
  — the list of transactions splits into separate time series
  — inside every time series we use the "nearest neighbour" approach
- we also used simple smoothing: a moving average with exponentially decreasing weights
  — this yields a slight improvement
- more advanced time series methods gave no improvement

Vicinities

- for small and for large strikes transactions are few
  — the previous transaction with the same strike may be far away in time
  — is not it better to take a more recent transaction from a neighbouring strike?
- let us consider vicinities of strikes
  — we predict implied volatility using the last transaction from a vicinity
- what is the right size for a vicinity?
  — we should use small vicinities in the middle and larger vicinities on the sides
- but how small and how large?
  — which neighbour is nearer to us, that in time or that in space?

Specialist Experts

- consider all sets of contiguous strikes $\{X_1, X_2, \ldots, X_k\}$, where k is a "diameter" parameter
- every vicinity specifies three experts:
  1. the expert working on transactions with strikes from this vicinity
  — when it sees a transaction with a strike from the vicinity, it outputs volatility from the previous transaction from the vicinity
  — when it sees a transaction with a strike from outside the vicinity, it sleeps
  2. the expert working on transactions with call options with strikes from this vicinity
  3. the expert working on transactions with put options with strikes from this vicinity

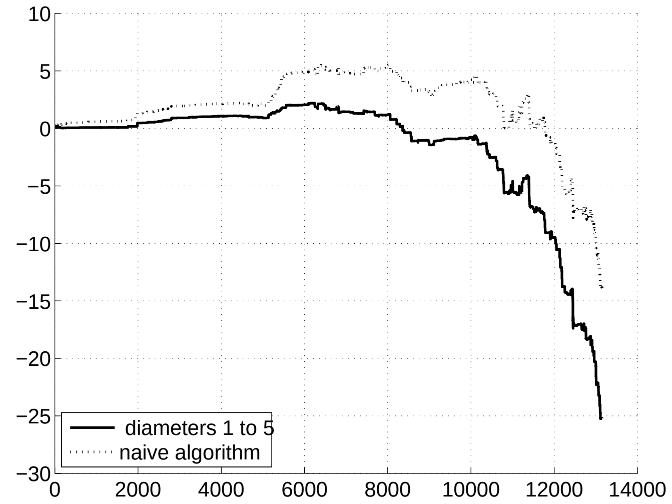
Results

- the following options were considered:
  1. Options on shares of Russian Energy Systems maturing in December 2006, 13K transactions
  2. Options on shares of Gazprom maturing in March 2007, 11K transactions
  3. Options on the RTSSE index (index is a portfolio of a special type) maturing in March 2007, 8.5K transactions
- we plot the adjusted loss

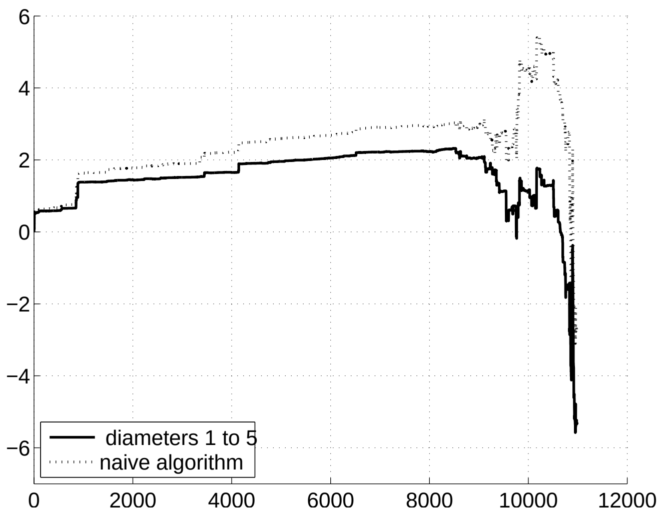
$$\text{Loss}(T) - \text{Loss}_{\text{RTSSE}}(T)$$

where $\text{Loss}_{\text{RTSSE}}(T)$ is the loss of a proprietary strategy (based on a parametric approximation for $\sigma(X)$)

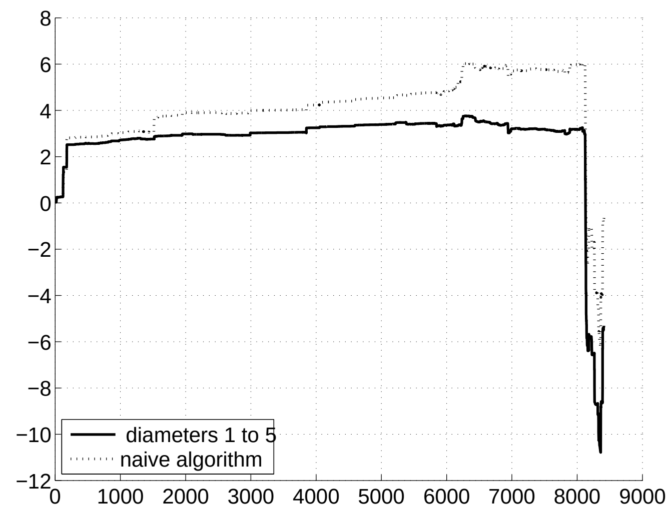
Russian Energy Systems



Gazprom



RTSSE Index



Discussion (1)

- results are comparable to those of a sliding window regression
  — regression is a standard way to model implied volatility
- as the maximum allowed diameter increases, the loss drops and then starts slowly going up
  — the regret is proportional to ln(number of experts) and grows very slowly
  — taking to many experts is rarely a problem: the algorithm will converge on the right ones

Discussion (2)

- vicinities of diameter 5 *alone* produce a poor result; but adding them to vicinities of sizes 1 to 4 improves the result.
  — even poor predictors work well somewhere

Dataset

- a kaggle challenge "What do we know"
- 4.851.475 examples
- each example is a record of a student answering a question
  — a log from some system training students for ACT, GMAT, and SAT

Data Fields

- timestamp
- student id
- question id
- track, subtrack, tags
  — what kind of question it is
- outcome: question answered or not

Game

- outcome space $\Omega = \{0, 1\}$
- prediction space $\Gamma = [0, 1]$
  — probability of the student answering correctly
- loss function: capped logarithmic loss ($\log_{10}$)
  — we can think that predictions are truncated to $[0.01, 0.99]$

Batch Setup

- all students who answered more than 6 questions are taken
- for each student the stream is cut at a random point > 6; the last record becomes a test example and all later records are discarded
- all remaining records form the kaggle training set
  — we have access to the future (but not for the same student)

Benchmark Method

- for each student there is a parameter α_i (meaning: student's strength)
- for each question there is a parameter β_j (meaning: question difficulty)

$$\Pr(\text{correct answer}) = \frac{e^{\alpha_i - \beta_j}}{1 + e^{\alpha_i - \beta_j}}$$

- the parameters α_i, β_j are fitted on the training set
- this is known as the Rasch model

Batch Results

- Rasch model mean loss on the kaggle test set is 0.2566
- the leader in kaggle competition achieves 0.2452
- for comparison $\log_{10} 2 = 0.3010$

On-line Mode

- we read the dataset example by example predicting the next outcome as we go along
  — we can only do this on the kaggle training set
- no access to the future

Constant Experts

- take a grid $\{0.1, 0.2, \ldots, 0.9\}$ of s numbers
- for every *student* take s experts making constant predictions for this student (irrespective of questions)
  — interpretation: each expert takes a view regarding the student's strength
- for every *question* take s experts making constant predictions for this question (irrespective of students)
  — interpretation: each expert takes a view regarding the question difficulty
- we take a uniform prior

Trofimov-Krichevsky

- for large s ($s = 9$ is OK) this is very similar to having a Trofimov-Krichevsky predictor for each student and for each each question
- a T-K is the Bayesian estimate for the probability of success in the Bernoulli model
  — a T-K predictor for student i predicts

$$\frac{\text{the number of correct answers the student has made so far} + 1}{\text{the number of questions the student has answered so far} + 2}$$

Results

- on the training set aggregating algorithm with constant experts suffers mean loss per element 0.2532
- on the kaggle test set the mean loss per element is 0.2717
- recall that the Rasch model loss on the kaggle test set is 0.2566

A comparison

- let us pick a retrospectively best constant expert for each question ("true difficulty")
  — the mean loss per question is 0.2631
- now let us pick a retrospectively best expert for each student ("true strength")
  — the mean loss per element is 0.2736
- and the loss of the aggregating algorithm is 0.2533

Tag Experts

- a tag expert watches a particular tag; when it occurs, the expert uses the T-K predictor on the current student predictions on questions with this tag
- a tag expert takes a view that its tag is informative; it represents a particular topic or subject and each student has a fixed strength in that subject
- aggregating algorithm with tag experts is generally inferior but...

Loss Structure

- this picture shows loss per element for students who answered more than a particular number of questions
  — the algorithm was still run on the whole dataset; the lower bound on the length is for reporting only
  — at the end averaging includes a small number of students