Talk 1: Predicting the unpredictable

Dean Foster (Amazon)

August 22, 2022

Plan for the next three days

- Goal: Getting comfortable with worst case data
  - I'll give a different proof sketch every session
  - I'll develop more philosophy than mechanics
  - Now "classic" (e.g. in COLT)
- Goal: Introducing calibration
  - First three lectures will be worst case sequential data
  - Last lecture will discuss traditional regression
- Goal: Fast methods for regression
  - Lecture 4 will discuss stepwise regression and other fast regression methods
  - This will set up the need for calibration in lecture 5

Schedule

	Monday	Tuesday	Wednesday
Morning (10-12)	Individual sequences	CMU speakers (Ann Lee, David Choi, Aaditya Ramdas)	Regression and big data
Lunch	students	students	students
Afternoon (2:30-4)	Calibeating	Macau and normal equations	Cross sectional Calibration
Dinner	faculty	faculty	

Asking the right question

Bad question: Can a forecaster guarantee the frequency of rain days to match their forecast?

time	0	1	2	3	4	5	6	7	8	$\cdots$
forecast	.4	.6	.7	.2	.6	.1	.6	.4	.4	$\cdots$
evil result	1	0	0	1	0	1	0	1	1	$\cdots$

Asking the right question

Bad question: Can a forecaster guarantee the frequency of rain days to match their forecast?

time	0	1	2	3	4	5	6	7	8	$\cdots$
forecast	.4	.6	.7	.2	.6	.1	.6	.4	.4	$\cdots$
evil result	1	0	0	1	0	1	0	1	1	$\cdots$

$$.4 \neq 1.0,\ .6 \neq 0,\ .7 \neq 0$$

Asking the right question

Bad question: Can a forecaster guarantee the frequency of rain days to match their forecast?

time	0	1	2	3	4	5	6	7	8	$\cdots$
forecast	.4	.6	.7	.2	.6	.1	.6	.4	.4	$\cdots$
evil result	1	0	0	1	0	1	0	1	1	$\cdots$

$$Y = \begin{cases} 0 & \text{if } \hat{y} > .5 \\ 1 & \text{if } \hat{y} \leq .5 \end{cases}$$

- For low forecasts, frequency is 1
- For high forecasts, frequency is 0
- Never calibrated (Oakes 1985)

Asking the right question

Bad question: Can a forecaster guarantee the frequency of rain days to match their forecast?

time	0	1	2	3	4	5	6	7	8	$\cdots$
forecast	.4	.6	.7	.2	.6	.1	.6	.4	.4	$\cdots$
evil result	1	0	0	1	0	1	0	1	1	$\cdots$

$$Y = \begin{cases} 0 & \text{if } \hat{y} > .5 \\ 1 & \text{if } \hat{y} \leq .5 \end{cases}$$

- For low forecasts, frequency is 1
- For high forecasts, frequency is 0
- Never calibrated (Oakes 1985)

So clearly the wrong question since nature wins instead of the statistician winning!

Choosing between two investments

- **Problem:** Suppose I have two friends who are hot-shot financial wizards. They come from different schools of thought and both believe the other to be totally clueless. So, in fact, I have one friend who is a financial wizard, and one friend who is an impostor. But, I don't know which is which!
- **Goal:** I want to get as rich as my financial wizard friend–whichever that empirically turns out to be.
- **No assumptions:** I will not make any probabilistic assumptions.

Setup for finance

- **Notation:**
  - A_t is the wealth of my first friend at time t
  - B_t is the wealth of my second friend
  - C_t is my wealth
  - w_t is fraction of my wealth A invests for me at time t
- $A_0 = B_0 = C_0 = 1$
- **Returns:**
  - $R_t^A = A_t / A_{t-1}$ is A's return
  - $R_t^B = B_t / B_{t-1}$ is B's return
  - $R_t^C = w_{t-1} R_t^A + (1 - w_{t-1}) R_t^B$ is my return
- **Goal:**

$$C_t \cong \max(A_t, B_t)$$

All three are growing "exponentially," so use $\log(C_t)$ instead. Now growing "linearly." So goal is:

$$\frac{\log(C_t)}{t} \cong \max(\frac{\log(A_t)}{t}, \frac{\log(B_t)}{t})$$

Invest with my best friend

- **Scheme:** Whichever friend is currently wealthier is "more likely" to be the financial wizard. So have her invest all my wealth:

$$w_t = \begin{cases} 1 & \text{if} \quad A_{t-1} \geq B_{t-1} \\ 0 & \text{if} \quad A_{t-1} < B_{t-1} \end{cases}$$

- **Evil data:**

time	0	1	2	3	4	5	6	7	8	⋯
A_t	1	1	2	2	4	4	8	8	16	⋯
B_t	1	2	2	4	4	8	8	16	16	⋯
w_t	1	0	1	0	1	0	1	0	1	⋯
C_t	1	1	1	1	1	1	1	1	1	⋯

- **growth rates:**
  - A's growth rate: $\ln(2)/2$
  - B's growth rate: $\ln(2)/2$
  - C's growth rate: 0

Equal weight

- **Scheme:** Always have each friend invest 1/2 of my wealth:

$$w_t = 1/2$$

- **Evil data:**

time	0	1	2	3	4	5	6	$\cdots$
A_t	1	2	4	8	16	32	64	$\cdots$
B_t	1	1	1	1	1	1	1	$\cdots$
w_t	$1/2$	$1/2$	$1/2$	$1/2$	$1/2$	$1/2$	$1/2$	$\cdots$
C_t	1	1.5	2.25	3.4	5.1	7.6	11.4	$\cdots$

- **growth rates:**
  - A's growth rate: $\ln(2)/2$
  - B's growth rate: 0
  - C's growth rate: $\ln(1.5)/2$

Value weighted

- **Scheme:** Have each invest in proportion to how well the have done so far.

$$w_t = \frac{A_{t-1}}{A_{t-1} + B_{t-1}}$$

- **No evil data exist!**
- **Growth rate of C:**

$$\frac{\ln(C_t)}{t} = \frac{\ln(A_t/2 + B_t/2)}{t} \geq \max\{\frac{\ln(A_t)}{t}, \frac{\ln(B_t)}{t}\} - \frac{\ln(2)}{t}$$

In particular:

$$\lim_{t\to\infty} \frac{\ln(C_t)}{t} - \max\{\frac{\ln(A_t)}{t}, \frac{\ln(B_t)}{t}\} = 0$$

Why did this work and not our first problem?

- Log smooths out differences
  - Our final wealth is only 1/2 of the better investor
- Intermediate value theorem avoids sharp edges
  - We don't have to pick a winner
  - We can hedge our bets by giving some to both

(ideas from Avrim Blum and Adam Kalai)

Choosing between two forecasts

- **Problem:** Suppose I have two friends who are hot-shot meteorologists. They come from different schools of thought and both believe the other to be totally clueless.
- **Goal:** I want to forecast the chance of rain as well as my hotshot friend.
- **No assumptions:** No probabilistic model.

Reduction to finance problem

We can reduce our problem to the finance one we solved already:

- Change money to probability
- Instead of multiplying returns, multiply probabilities
- Instead of log wealth, look at log probability loss

Reduction to finance problem

We can reduce our problem to the finance one we solved already:

- Change money to probability
- Instead of multiplying returns, multiply probabilities
- Instead of log wealth, look at log probability loss

Theorem

A Bayesian combination can do as well as the better meteorologist using log probability loss.

History

Proved for the “first time” in all the following papers:

- “Controlled Random Walks”
- “On Pseudo-games”
- “A Randomized Rule for Selecting Forecasts”
- “Approximating the Bayes risk in Repeated Plays”
- “Aggregating Strategies”
- “Compression of Individual Sequences via Variable-Rate Coding”
- “Universal Portfolios”

History

Proved for the “first time” in all the following papers:

- “Controlled Random Walks”
- “On Pseudo-games”
- “A Randomized Rule for Selecting Forecasts”
- “Approximating the Bayes risk in Repeated Plays”
- “Aggregating Strategies”
- “Compression of Individual Sequences via Variable-Rate Coding”
- “Universal Portfolios”

Real first time:

- Hannan (1955) proved and stated it
- Blackwell (1954 / 1990) proved but didn’t bother stating it until much later

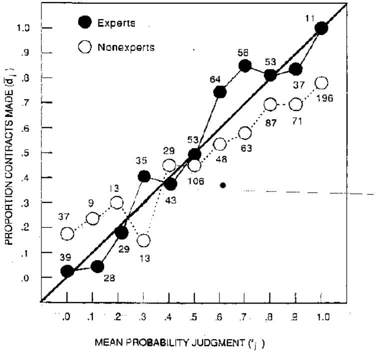
(F. and Vohra 1999 for more history)

Back to calibration

How can we make the calibration game winnable?

- One smooth scoring rule instead of each forecast sold separately
- Allow randomization to get intermediate value to hold

Calibration: Smooth function



$\rho(p)$ = fraction of successes
$n(p)$ = number of forecasts
N = total

$$C = \text{calibration score} = \sum_p \frac{n(p)}{N} (\rho(p) - p)^2$$

Theorem (Oakes 1985)

Every deterministic forecast has a sequence with $C \geq 1/4$.

So we need randomization

We'll sneak up on calibration by reducing it to other problems:

- First we will introduce "no internal regret"
- We'll prove that exists by reducing it to regular regret
- Then we'll prove calibration by reducing it to no internal-regret

Regret

- Consider a set of k actions
- Let $U_t(i)$ be the utility of action i at time t
- Let $c_t \in \{1, \ldots, k\}$ pick among these k different actions
- Utility of this choice is: $U_t(c_t)$.

Regret

- Consider a set of k actions
- Let $U_t(i)$ be the utility of action i at time t
- Let $c_t \in \{1, \ldots, k\}$ pick among these k different actions
- Utility of this choice is: $U_t(c_t)$.

"We look before and after, And pine for what is not."

Percy Bysshe Shelley

Regret

- Consider a set of k actions
- Let $U_t(i)$ be the utility of action i at time t
- Let $c_t \in \{1, \ldots, k\}$ pick among these k different actions
- Utility of this choice is: $U_t(c_t)$.

"I wish I'd bet on red any time I had bet on black, and I wish I'd bet on black any time I had bet on red."

Winston Churchill

Regret

- Consider a set of k actions
- Let $U_t(i)$ be the utility of action i at time t
- Let $c_t \in \{1, \ldots, k\}$ pick among these k different actions
- Utility of this choice is: $U_t(c_t)$.

Definition

The regret generated by changing i's to j's is:

$$R_t^{i \to j} \equiv \sum_{s=1}^{t} I_{c_s = i} \left(U_s(j) - U_s(i) \right).$$

Definition

The regret generated by changing *i*'s to *j*'s is:

$$R_t^{i \to j} \equiv \sum_{s=1}^{t} I_{c_s = i} \left(U_s(j) - U_s(i) \right).$$

Definition

The regret generated by changing *i*'s to *j*'s is:

$$R_t^{i\to j} \equiv \sum_{s=1}^{t} I_{c_s=i} \left(U_s(j) - U_s(i)\right).$$

Definition (Traditional regret)

$$\max_j \sum_i R_t^{i\to j} = o(t).$$

Definition

The regret generated by changing *i*'s to *j*'s is:

$$R_t^{i \to j} \equiv \sum_{s=1}^{t} I_{c_s = i} \left(U_s(j) - U_s(i) \right).$$

Definition (Internal regret)

$$(\forall i, \forall j) \quad R_t^{i \to j} \leq o(t).$$

Definition

The regret generated by changing i's to j's is:

$$R_t^{i \to j} \equiv \sum_{s=1}^{t} I_{c_s = i} \left(U_s(j) - U_s(i) \right).$$

Definition (Swap regret)

$$\sum_i \max_j R_t^{i \to j} = o(t).$$

Definition

The regret generated by changing *i*'s to *j*'s is:

$$R_t^{i\to j} \equiv \sum_{s=1}^{t} I_{c_s=i} \left(U_s(j) - U_s(i)\right).$$

- Assume we have an algorithm with small regret.
- We'll use it to construct a no-internal regret algorithm

Definition

The regret generated by changing *i*'s to *j*'s is:

$$R_t^{i \to j} \equiv \sum_{s=1}^{t} I_{c_s = i} \left(U_s(j) - U_s(i) \right).$$

- For each action i have a no regret algorithm which runs whenever we play i
- This will pick good alternative actions
- So $R^{i \to j} \neq o(t)$.
- But we are not playing i anymore!

Definition

The regret generated by changing *i*'s to *j*'s is:

$$R_t^{i \to j} \equiv \sum_{s=1}^{t} I_{c_s=i} \left(U_s(j) - U_s(i) \right).$$

Let P^{ji} be the probability of algorithm *i* playing *j*

$$\pi = P\pi$$

Now we play *j* as suggested by algorithm *i*.

Definition

The regret generated by changing *i*'s to *j*'s is:

$$R_t^{i \to j} \equiv \sum_{s=1}^{t} I_{c_s = i} \left(U_s(j) - U_s(i) \right).$$

Let P^{ji} be the probability of algorithm i playing j

$$\pi = P\pi$$

Now we play j as suggested by algorithm i.

- We called this a flow condition
- Sergiu Hart called it regret matching
- Blum and Mansour didn't bother to name it!

Definition

The regret generated by changing *i*'s to *j*'s is:

$$R_t^{i \to j} \equiv \sum_{s=1}^{t} I_{c_s = i} \left(U_s(j) - U_s(i) \right).$$

Let P^{ji} be the probability of algorithm i playing j

$$\pi = P\pi$$

Now we play j as suggested by algorithm i.

Theorem (Foster & Vohra 1995)

There exist a no internal regret algorithm, namely $\max_{ij} R_t^{i \to j} = o(t)$.

Theorem

Swap regret $\Leftrightarrow$ *Internal regret* $\Rightarrow$ *traditional regret.*

Theorem

Swap regret ⇔ Internal regret ⇒ traditional regret.

Proof:

$$\underbrace{k \max_{ij} R_t^{i \to j}}_{internal} \geq \underbrace{\sum_i \max_j R_t^{i \to j}}_{swap} \geq \underbrace{\max_{ij} R_t^{i \to j}}_{internal} \geq (1/k) \underbrace{\max_j \sum_i R_t^{i \to j}}_{traditional}.$$

Calibration

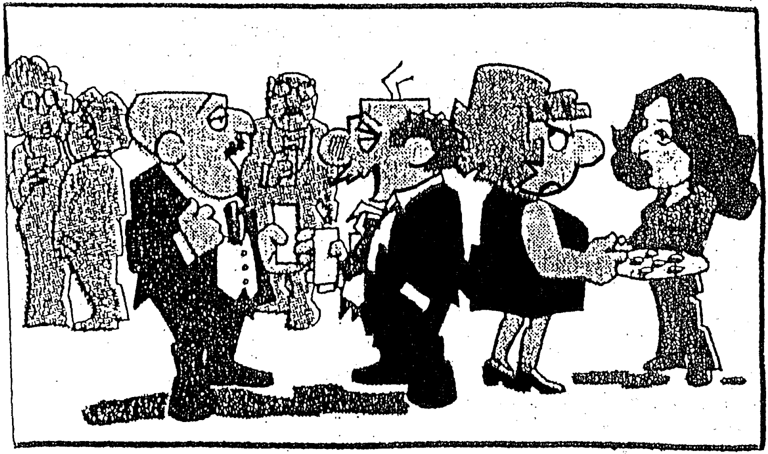
Consider the k actions: $\widehat{p}_t \in \{0, \frac{1}{k}, \frac{2}{k}, \ldots, \frac{k-1}{k}, 1\}$.

- Use quadratic loss
- If the frequency is far from $\widehat{p}_t$ then switching to a $\frac{i}{k}$ which is closer
- Running this no-internal regret algorithm will generate a calibrated forecast.

Theorem (Foster & Vohra 1991-1998)

There exist a randomized calibrated forecast.

What is an equilibrium?



"LORETTA'S DRIVING BECAUSE I'M DRINKING, AND I'M DRINKING BECAUSE SHE'S DRIVING."

Definition of Nash equilibrium

Definition (Nash Equilibrium)

For two players, with sigma fields $\mathcal{F}$ and $\mathcal{G}$, and utilities U_f and U_g then f playing X and g playing Y is a Nash equilibrium if:

- $E(U_f(X, Y)|\mathcal{F}) \geq E(U_f(x, Y)|\mathcal{F})$ for all x
- $E(U_g(X, Y)|\mathcal{G}) \geq E(U_g(X, y)|\mathcal{G})$ for all y
- $\mathcal{F}$ is independent of $\mathcal{G}$.

Definition of Nash equilibrium

Definition (~~Nash~~ Equilibrium)

For two players, with sigma fields $\mathcal{F}$ and $\mathcal{G}$, and utilities U_f and U_g then f playing X and g playing Y is a Nash equilibrium if:

- $E(U_f(X, Y)|\mathcal{F}) \geq E(U_f(x, Y)|\mathcal{F})$ for all x
- $E(U_g(X, Y)|\mathcal{G}) \geq E(U_g(X, y)|\mathcal{G})$ for all y
- ~~$\mathcal{F}$ is independent of $\mathcal{G}$.~~

Called a Correlated equilibrium.

Definition of Nash equilibrium

Definition (Correlated Equilibrium)

For two players, with sigma fields $\mathcal{F}$ and $\mathcal{G}$, and utilities U_f and U_g then f playing X and g playing Y is a Nash equilibrium if:

- $E(U_f(X, Y)|\mathcal{F}) \geq E(U_f(x, Y)|\mathcal{F})$ for all x
- $E(U_g(X, Y)|\mathcal{G}) \geq E(U_g(X, y)|\mathcal{G})$ for all y

Roger Myerson: "2 out of 3 intelligent species discover Correlated equilibrium before Nash equilibrium."

I've been quote Roger before he got his Nobel in 2007.

Fictitious play model

- The first player predicts the second player
- The second player predicts the first player
- Each plays a best reply to their predictions
- Called fictitious play

Fictitious play model

- The first player predicts the second player
- The second player predicts the first player
- Each plays a best reply to their predictions
- Called fictitious play

Theorem (Foster and Vohra, 1998)

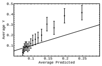
If both players use a calibrated forecast, they converge to a Correlated equilibrium.

Where to next?

- This afternoon we'll remove some of the randomness
  - Allows convergence to Nash equilibrium
- We'll see how to be calibrated AND find patterns in data
  - Called “calibeating”
- See you after lunch!

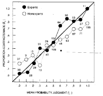
Summary handout

Calibration



- Works well for big data since only costs a few more degrees of freedom.
- "Variable selection in data mining: Building a predictive model for bankruptcy," Foster and Stine, *JASA*, 2004.
- "Efficient Learning of Generalized Linear and Single Index Models with Isotonic Regression," Kakade, Kalai, Kanade, and Shamir, 2011.

- "Precision and Accuracy of Judgmental Estimation," Foster and Yaniv, *Journal of Behavioral Decision Making* (1997).
- "Graininess of Judgment Under Uncertainty: An Accuracy - informativeness Tradeoff," Foster and Yaniv *Journal of Experimental Psychology: General*, 1995.
- We looked at confidence intervals.
- Humans actually are responding to the social utility function.



"Suppose in a long (conceptually infinite) sequence of weather forecasts, we look at all those days for which the forecast probability of precipitation was, say, close to some given value p and then determine the long run proportion f of such days on which the forecast event (rain) in fact occurred. If $f = p$ the forecaster may be termed well calibrated."

Phillip Dawid

- "Asymptotic Calibration," Foster and Vohra, *Biometrika*, 1998, (also Foster *GEB* 1999).
- "Regret in the On-line Decision Problem," Foster and Vohra, *GEB* 1999. (See also *AI-STATS* 2012 and *MOR* 2014.)
- "Deterministic Calibration and Nash Equilibrium" Foster and Kakade, *COLT*, 2004, (see also Foster & Hart, *JPE* 2021.)

Games

- "Calibrated Learning and Correlated Equilibrium," Foster and Vohra *Games and Economic Behavior*, 1997.
  - Playing calibrated forecasts will lead to correlated equilibria
  - Playing no-interal regret actions will converge to correlated equilibria
- Extended in "A general class of adaptive strategies, " by Hart and Mas-Colell 2001.

"If there is intelligent life on other planets, in a majority of them, they would have discovered correlated equilibrium before Nash equilibrium."

Roger Myerson

Nash is hard

- Yes: You can learn NE from a grain of truth. (Kalai and Lehrer 1993).
- No: Not exactly. (Nachbar 1997, Foster and Young 2001)
- Yes: Via exhaustive search–i.e. very slowly. (Foster and Young 2006)
- No: Hart and Mas-Colell 2011.
- Yes: Via public, deterministic calibration which is very slow (Foster and Kakade 2008, Foster and Hart 2018)
- For all but the smallest games, it is basically no.

Recommendations

- Use isotonic link functions to calibrate regressions (Wednesday)
- Use fixed point based calibration for time series (Tuesday)
- Use no-internal regret for game theory (Monday morning)
- Let go of Nash equilibrium (Monday morning)

References

- Blackwell, David (1956) “An Analog of the Minimax Theorem for Vector Payoffs,” *Pacific Journal of Mathematics,* **6**.
- Blackwell, David (1990) “The Prediction of Sequences, ” mimeo.
- Cover, Thomas (1991) “Universal Portfolios,” *Mathematical Finance*.
- Foster (1991) “A worst Case Analysis of Prediction,” *Annals of Statistics,* **19**, 1084-1090.
- Vovk (1991) “Aggregating Strategies,” *Proceedings of the 3rd annual workshop on Computational Learning Theory*, 1991, 371 - 383.
- Foster and Vohra (1993) “A Randomization Rule for Selecting Forecasts,” *Operations Research*, 704-709.
- Foster and Vohra (1995) “Asymptotic Calibration.”