data science @ NYT

Chris Wiggins

Aug 8/9, 2016

Contents

0.1 Outline . . . 3
0.2 0. Thank the organizers! . . . 3

1 Lecture 1: overview of ds@NYT 3

2 Lecture 2: predictive modeling @ NYT 3
2.1 desc/pred/pres . . . 3
2.2 blossom example . . . 3
2.3 blossom + boosting ('exponential') . . . 3
2.4 tangent: logistic function as surrogate loss function . . . 3
2.5 boosting 1 . . . 6
2.6 boosting 1 . . . 6
2.7 predicting people . . . 6
2.8 predicting people (reminder) . . . 6
2.9 single copy (reminder) . . . 6
2.10 example in CAR (computer assisted reporting) . . . 6
2.11 example in CAR (computer assisted reporting) . . . 6
2.12 computer assisted reporting . . . 9

3 Lecture 3: prescriptive modeling @ NYT 9
3.1 the natural abstraction . . . 9
3.2 maximizing outcome . . . 9
3.3 examples . . . 10
3.4 reinforcement/machine learning/graphical models . . . 10
3.5 in general . . . 10
3.6 randomized controlled trial . . . 10
3.7 POISE: calculation, estimation, optimization . . . 11
3.8 POISE setup and Goal . . . 11
3.9 POISE math: IS+Monte Carlo estimation=ISE . . . 11
3.10 Observation (cf. Bottou) . . . 12
3.11 Reduction (cf. Langford„ ('05, '08, '09)) . . . 12
3.12 Reduction w/optimistic complication . . . 12

3.13 POISE punchlines . . . 12
3.14 tangent: causality as told by an economist . . . 13
3.15 Q-note: "generalizing" Monte Carlo w/kernels . . . 13
3.16 Q-note: application w/strata+matching, setup . . . 13
3.17 Q-note: application w/strata+matching, payoff . . . 13
3.18 causality, as understood in marketing . . . 14
3.19 unobserved confounders vs. "causality" modeling . . . 14
3.20 cautionary tale problem: Simpson's paradox . . . 15
3.21 confounded approach: quasi-experiments + instruments . . . 15
3.22 IV: graphical model assumption . . . 15
3.23 IV: graphical model assumption (sideways) . . . 16
3.24 IV: review s/OLS/MOM/ (E is empirical average) . . . 16
3.25 IV: binary, binary case (aka "Wald estimator") . . . 16
3.26 bandits: obligatory slide . . . 17
3.27 bandits . . . 17
3.28 TS: connecting w/"generative causal modeling" 0 . . . 17
3.29 TS: connecting w/"generative causal modeling" 1 . . . 18
3.30 TS: connecting w/"generative causal modeling" 2 . . . 18
3.31 That sounds hard. . . . 19
3.32 Thompson sampling: results (2011) . . . 19
3.33 TS: words . . . 20
3.34 TS: p-code . . . 20
3.35 TS: Bernoulli bandit p-code . . . 20
3.36 TS: Bernoulli bandit p-code (results) . . . 20
3.37 UCB1 (2002), p-code . . . 20
3.38 TS: with context . . . 20
3.39 LinUCB: UCB with context . . . 20
3.40 TS: with context (results) . . . 25
3.41 Bandits: Regret via Lai and Robbins (1985) . . . 25
3.42 Thompson sampling (1933) and optimality (2013) . . . 25
3.43 other 'Causalities': structure learning . . . 25
3.44 other 'Causalities': potential outcomes . . . 25

4 Lecture 4: descriptive modeling @ NYT 26
4.1 review: (latent) inference and clustering . . . 26
4.2 actual math . . . 26
4.3 tangent: more math on GMMs, part 1 . . . 27
4.4 tangent: more math on GMMs, part 2 . . . 27
4.5 tangent: Gaussians $\in$ exponential family . . . 27
4.6 tangent: variational joy $\in$ exponential family . . . 28
4.7 clustering and inference: GMM/k-means case study . . . 28
4.8 general framework: E+M/variational . . . 28
4.9 example application: LDA+topics . . . 31
4.10 rec engine via CTM . . . 31
4.11 recall: recommendation via factoring . . . 31
4.12 CTM: combined loss function . . . 31

4.13 CTM: updates for factors 31
4.14 CTM: (via Jensen's, again) bound on loss 31

5 Lecture 5 data product **31**
5.1 data science and design thinking 31
5.2 Thanks! 31

0.1 Outline

1. overview of DS@NYT
2. prediction + supervised learning
3. prescription, causality, and RL
4. description + inference
5. (if interest) designing data products

0.2 0. Thank the organizers!

1 Lecture 1: overview of ds@NYT

2 Lecture 2: predictive modeling @ NYT

2.1 desc/pred/pres

- caveat: difference between observation and experiment. why?

2.2 blossom example

2.3 blossom + boosting ('exponential')

2.4 tangent: logistic function as surrogate loss function

- define $f(x) \equiv \log p(y=1|x)/p(y=-1|x) \in R$
- $p(y=1|x) + p(y=-1|x) = 1 \rightarrow p(y|x) = 1/(1+\exp(-yf))$
- $-\log_2 p(\{y\}_1^N) = \sum_i \log_2 \left(1 + \mathrm{e}^{-y_i f(x_i)}\right) \equiv \sum_i \ell(y_i f(x_i))$
- $\ell'' > 0$, $\ell(\mu) > 1[\mu < 0]$ $\forall \mu \in R$.
- $\therefore$ maximizing log-likelihood is minimizing a surrogate convex loss function for classification (though not strongly convex, cf. Yoram's talk)
- but $\sum_i \log_2 \left(1 + \mathrm{e}^{-y_i w^T h(x_i)}\right)$ not as easy as $\sum_i \mathrm{e}^{-y_i w^T h(x_i)}$

Figure 1: prepping slides until last minute

descriptive:	specify x; learn $z(x)$ or $p(z\|x)$ where z is "simpler" than x
predictive:	specify x and y; learn to predict y from x
prescriptive:	specify x, y, and a; learn to prescribe a given x to maximize y

Figure 2: desc/pred/pres

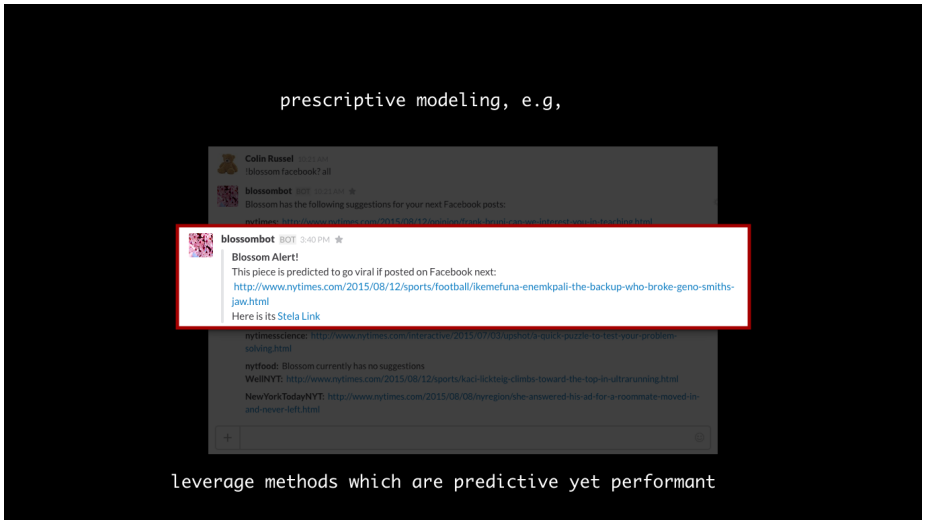


Figure 3: Reminder: Blossom

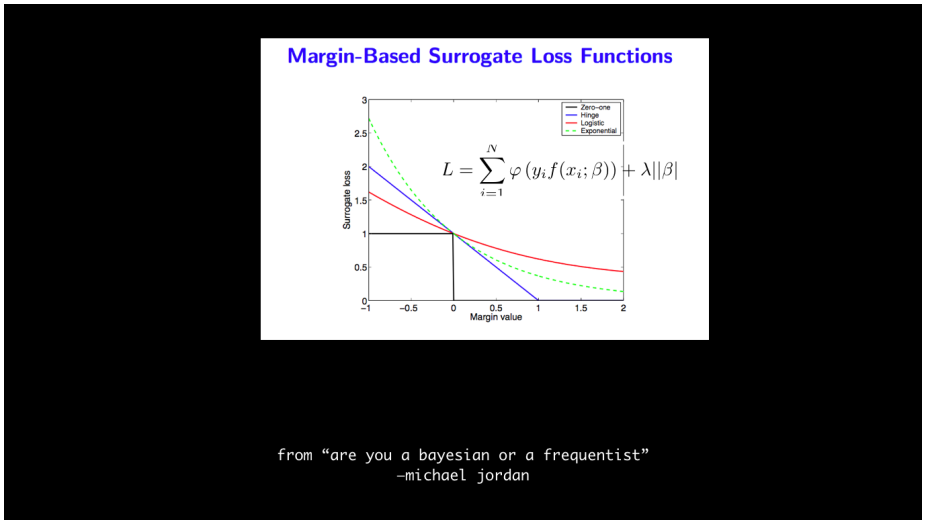


Figure 4: Reminder: Surrogate Loss Functions

2.5 boosting 1

L exponential surrogate loss function, summed over examples:

- $L[F] = \sum_i \exp\left(-y_i F(x_i)\right)$
- $= \sum_i \exp\left(-y_i \sum_{t'}^{t} w_{t'} h_{t'}(x_i)\right) \equiv L_t(\mathbf{w}_t)$
- Draw $h_t \in \mathcal{H}$ large space of rules s.t. $h(x) \in \{-1, +1\}$
- label $y \in \{-1, +1\}$

2.6 boosting 1

L exponential surrogate loss function, summed over examples:

- $L_{t+1}(\mathbf{w}_t; w) \equiv \sum_i d_i^t \exp\left(-y_i w h_{t+1}(x_i)\right)$
- $= \sum_{y=h'} d_i^t \mathrm{e}^{-w} + \sum_{y \neq h'} d_i^t \mathrm{e}^{+w} \equiv \mathrm{e}^{-w} D_+ + \mathrm{e}^{+w} D_-$
- $\therefore w_{t+1} = \mathrm{argmin}_w L_{t+1}(w) = (1/2) \log D_+ / D_-$
- $L_{t+1}(\mathbf{w}_{t+1}) = 2\sqrt{D_+ D_-} = 2\sqrt{\nu_+(1-\nu_+)}/D$, where $0 \leq \nu_+ \equiv D_+/D = D_+/L_t \leq 1$
- update example weights $d_i^{t+1} = d_i^t \mathrm{e}^{\mp w}$

Punchlines: sparse, predictive, interpretable, fast (to execute), and easy to extend, e.g., trees, flexible hypotheses spaces, L_1, L_∞[1], ...

2.7 predicting people

- "customer journey" prediction
  - fun covariates
  - observational complication v structural models

2.8 predicting people (reminder)

2.9 single copy (reminder)

2.10 example in CAR (computer assisted reporting)

2.11 example in CAR (computer assisted reporting)

- cf. Friedman's "Statistical models and Shoe Leather"[2]
- Takata airbag fatalities
- 2219 labeled[3] examples from 33,204 comments

[1] Duchi + Singer "Boosting with structural sparsity" ICML '09

[2] Freedman, David A. "Statistical models and shoe leather." Sociological methodology 21.2 (1991): 291-313.

[3] By Hiroko Tabuchi, a Pulitzer winner

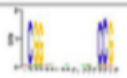
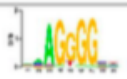
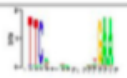
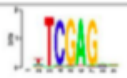
TFNAME	DB-MOTIF	MOTIF	DBNAME	d(p,q)
CBF1	CACGTG		YPD	0.032635
CGG everted repeat	CGGN*CCG		YPD	0.032821
MSN2			TRANSFAC	0.085626
HSF1	TTCNNNGAA		SCPD	0.102410
XBP1			TRANSFAC	0.140561

Figure 5: both in science and in real world, feature analysis guides future experiments

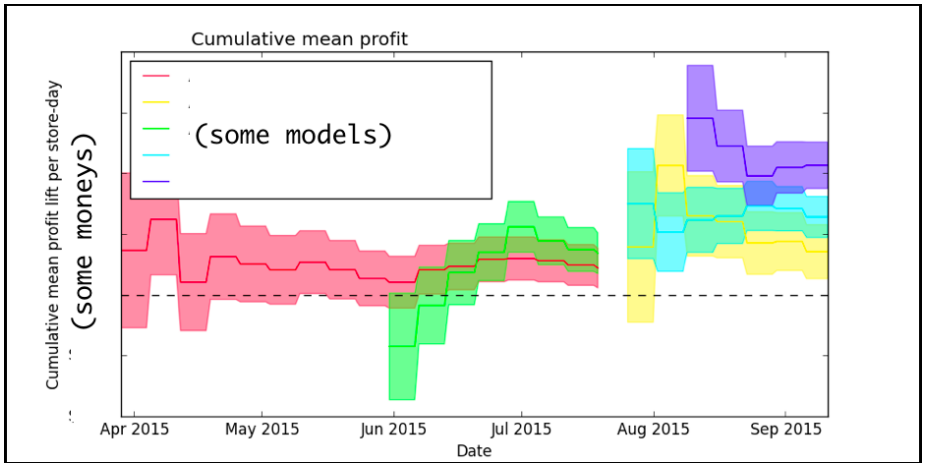


Figure 6: from Lecture 1

Figure 7: Tabuchi article

- cf. Box's "Science and Statistics"[4]

2.12 computer assisted reporting

- Impact

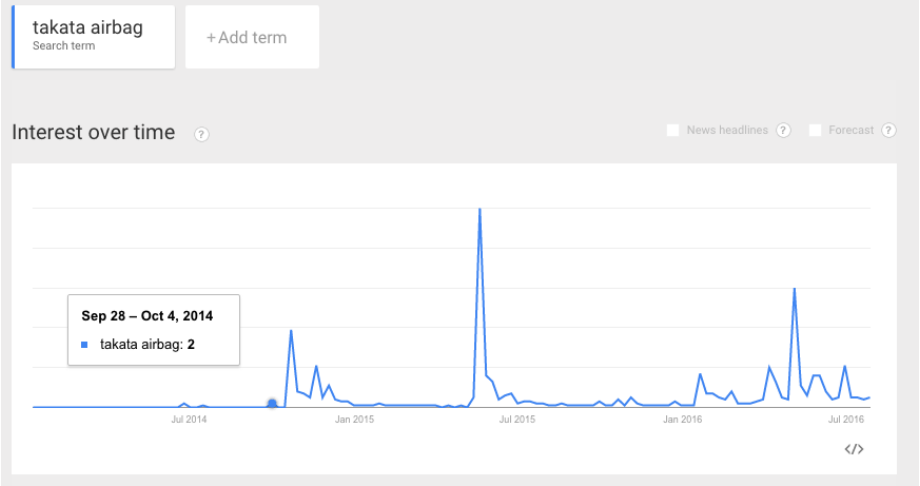


Figure 8: impact

3 Lecture 3: prescriptive modeling @ NYT

3.1 the natural abstraction

- operators[5] make decisions
- faster horses v. cars
- general insights v. optimal policies

3.2 maximizing outcome

- the problem: maximizing an outcome over policies...
- ...while inferring causality from observation
- different from predicting outcome in absence of action/policy

[4] Science and Statistics, George E. P. Box Journal of the American Statistical Association, Vol. 71, No. 356. (Dec., 1976), pp. 791-799.

[5] In the sense of business deciders; that said, doctors, including those who operate, also have to make decisions, cf., personalized medicines

3.3 examples

- observation is not experiment
  - e.g., (Med.) smoking hurts vs unhealthy people smoke
  - e.g., (Med.) affluent get prescribed different meds/treatment
  - e.g., (life) veterans earn less vs the rich serve less[6]
  - e.g., (life) admitted to school vs learn at school?

3.4 reinforcement/machine learning/graphical models

- key idea: model joint $p(y, a, x)$
- explore/exploit: family of joints $p_\alpha(y, a, x)$
- "causality": $p_\alpha(y, a, x) = p(y|a, x)p_\alpha(a|x)p(x)$ "a causes y"
- nomenclature: 'response', 'policy'/'bias', 'prior' above

3.5 in general

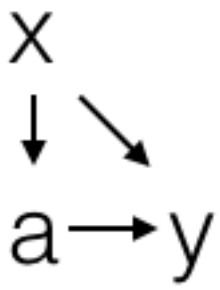


Figure 9: policy/bias, response, and prior define the distribution

also describes both the 'exploration' and 'exploitation' distributions

3.6 randomized controlled trial

also Pearl's 'do' distribution: a distribution with "no arrows" pointing to the action variable.

[6] Angrist, Joshua D. (1990). "Lifetime Earnings and the Vietnam Draft Lottery: Evidence from Social Security Administrative Records". American Economic Review 80 (3): 313–336.

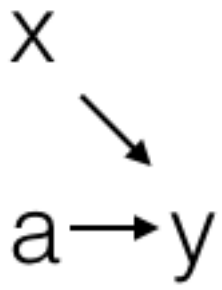


Figure 10: RCT: 'bias' removed, random 'policy' (response and prior unaffected)

3.7 POISE: calculation, estimation, optimization

- POISE: "policy optimization via importance sample estimation"
- Monte Carlo importance sampling estimation
  - aka "off policy estimation"
  - role of "IPW"
- reduction
- normalization
- hyper-parameter searching
- unexpected connection: personalized medicine

3.8 POISE setup and Goal

- "a causes y" $\iff \exists$ family $p_\alpha(y,a,x) = p(y|a,x)p_\alpha(a|x)p(x)$
- define off-policy/exploration distribution $p_-(y,a,x) = p(y|a,x)p_-(a|x)p(x)$
- define exploitation distribution $p_+(y,a,x) = p(y|a,x)p_+(a|x)p(x)$
- Goal: Maximize $E_+(Y)$ over $p_+(a|x)$ using data drawn from $p_-(y,a,x)$.

. . .

notation: $\{x,a,y\} \in \{X,A,Y\}$ i.e., $E_\alpha(Y)$ is not a function of y

3.9 POISE math: IS+Monte Carlo estimation=ISE

i.e, "importance sampling estimation"

- $E_+(Y) \equiv \sum_{yax} yp_+(y,a,x)$
- $E_+(Y) = \sum_{yax} yp_-(y,a,x)(p_+(y,a,x)/p_-(y,a,x))$
- $E_+(Y) = \sum_{yax} yp_-(y,a,x)(p_+(a|x)/p_-(a|x))$
- $E_+(Y) \approx N^{-1} \sum_i y_i(p_+(a_i|x_i)/p_-(a_i|x_i))$

. . .

let's spend some time getting to know this last equation, the importance sampling estimate of outcome in a "causal model" ("a causes y") among $\{y, a, x\}$

3.10 Observation (cf. Bottou[7])

- factorizing $P_{\pm}(x)$: $\frac{P_+(x)}{P_-(x)} = \Pi_{\text{factors}} \frac{P_{+\text{but not}-}(x)}{P_{-\text{but not}+}(x)}$
- origin: importance sampling $E_q(f) = E_p(fq/p)$ (as in variational methods)
- the "causal" model $p_\alpha(y, a, x) = p(y|a, x)p_\alpha(a|x)p(x)$ helps here
- factors left over are numerator ($p_+(a|x)$, to optimize) and denominator ($p_-(a|x)$, to infer if not a RCT)
- unobserved confounders will confound us (later)

3.11 Reduction (cf. Langford[8],[9],[10] ('05, '08, '09))

- consider numerator for deterministic policy: $p_+(a|x) = 1[a = h(x)]$
- $E_+(Y) \propto \sum_i (y_i/p_-(a|x))1[a = h(x)] \equiv \sum_i w_i 1[a = h(x)]$
- Note: $1[c = d] = 1 - 1[c \neq d]$
- $\therefore E_+(Y) \propto \text{constant} - \sum_i w_i 1[a \neq h(x)]$
- $\therefore$ reduces policy optimization to (weighted) classification

3.12 Reduction w/optimistic complication

- Prescription $\iff$ classification $L = \sum_i w_i 1[a_i \neq h(x_i)]$
- weight $w_i = y_i/p_-(a_i|x_i)$, inferred or RCT
- destroys measure by treating $p_-(a|x)$ differently than $1/p_-(a|x)$
- normalize as $\tilde{L} \equiv \frac{\sum_i y1[a_i \neq h(x_i)]/p_-(a_i|x_i)}{\sum_i 1[a_i \neq h(x_i)]/p_-(a_i|x_i)}$
- destroys lovely reduction
- simply[11] $L(\lambda) = \sum_i (y_i - \lambda)1[a_i \neq h(x_i)]/p_-(a_i|x_i)$
- hidden here is a 2nd parameter, in classification, $\therefore$ harder search

3.13 POISE punchlines

- allows policy planning even with implicit logged exploration data[12]

[7] Counterfactual Reasoning and Learning Systems, arXiv:1209.2355

[8] Langford & Zadrozny "Relating Reinforcement Learning Performance to Classification Performance" ICML 2005

[9] Beygelzimer & Langford "The offset tree for learning with partial labels" (KDD 2009)

[10] Tutorial on "Reductions" (including at ICML 2009)

[11] Suggestion by Dan Hsu

[12] Strehl, Alex, et al. "Learning from logged implicit exploration data." Advances in Neural Information Processing Systems. 2010.

- e.g., two hospital story
- "personalized medicine" is also a policy
- abundant data available, under-explored IMHO

3.14 tangent: causality as told by an economist

different, related goal

- they think in terms of ATE/ITE instead of policy
  - ATE
    - $\tau \equiv E_0(Y|a=1) - E_0(Y|a=0) \equiv Q(a=1) - Q(a=0)$
  - CATE aka Individualized Treatment Effect (ITE)
    - $\tau(x) \equiv E_0(Y|a=1,x) - E_0(Y|a=0,x)$
    - $\equiv Q(a=1,x) - Q(a=0,x)$

3.15 Q-note: "generalizing" Monte Carlo w/kernels

- *MC*: $E_p(f) = \sum_x p(x)f(x) \approx N^{-1} \sum_{i \sim p} f(x_i)$
- K: $p \approx N^{-1} \sum_i K(x|x_i)$
- $\Rightarrow \sum_x p(x)f(x) \approx N^{-1} \sum_i \sum_x f(x)K(x|x_i)$
- K can be any normalized function, e.g., $K(x|x_i) = \delta_{x,x_i}$, which yields *MC*.
- multivariate $E_p(f) \approx N^{-1} \sum_i \sum_{yax} f(y,a,x)K_1(y|y_i)K_2(a|a_i)K_3(x|x_i)$

3.16 Q-note: application w/strata+matching, setup

Helps think about economists' approach:

- $Q(a,x) \equiv E(Y|a,x) = \sum_y y p(y|a,x) = \sum_y y \frac{p_-(y,a,x)}{p_-(a|x)p(x)}$
- $= \frac{1}{p_-(a|x)p(x)} \sum_y y p_-(y,a,x)$
- stratify x using $z(x)$ such that $\cup z = X$, and $\cap z, z' = \varnothing$
- $n(x) = \sum_i 1[z(x_i) = z(x)]$=number of points in x's stratum
- $\Omega(x) = \sum_{x'} 1[z(x') = z(x)]$=area of x's stratum
- $\therefore K_3(x|x_i) = 1[z(x) = z(x_i)]/\Omega(x)$
- as in *MC*, $K_1(y|y_i) = \delta_{y,y_i}$, $K_2(a|a_i) = \delta_{a,a_i}$

3.17 Q-note: application w/strata+matching, payoff

- $\sum_y y p_-(y,a,x) \approx N^{-1}\Omega(x)^{-1} \sum_{a_i=a, z(x_i)=z(x)} y_i$
- $p(x) \approx (n(x)/N)\Omega(x)^{-1}$
- $\therefore Q(a,x) \approx p_-(a|x)^{-1} n(x)^{-1} \sum_{a_i=a, z(x_i)=z(x)} y_i$

. . .

"matching" means: choose each z to contain 1 positive example & 1 negative example,

- $p_-(a|x) \approx 1/2, n(x) = 2$
- $\therefore \tau(a,x) = Q(a=1,x) - Q(a=0,x) = y_1(x) - y_0(x)$
- z-generalizations: graphs, digraphs, k-NN, "matching"
- K-generalizations: continuous a, any metric or similarity you like,...

. . .

IMHO underexplored

3.18 causality, as understood in marketing

- a/b testing and RCT
- yield optimization
- Lorenz curve (vs ROC plots)

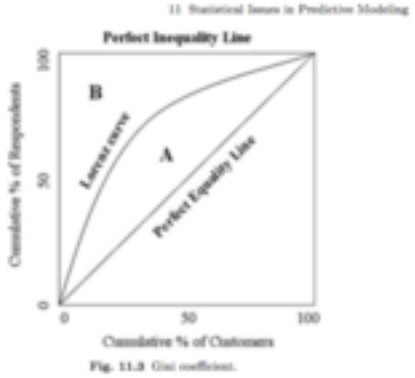


Figure 11: Blattberg, Robert C., Byung-Do Kim, and Scott A. Neslin. Database Marketing, Springer New York, 2008

3.19 unobserved confounders vs. "causality" modeling

- truth: $p_\alpha(y,a,x,u) = p(y|a,x,u)p_\alpha(a|x,u)p(x,u)$
- but: $p_+(y,a,x,u) = p(y|a,x,u)p_-(a|x)p(x,u)$
- $E_+(Y) \equiv \sum_{yaxu} yp_+(yaxu) \approx N^{-1} \sum_{i \sim p_-} y_i p_+(a|x)/p_-(a|x,u)$
- denominator can not be inferred, ignore at your peril

3.20 cautionary tale problem: Simpson's paradox

- a: admissions (a=1: admitted, a=0: declined)
- x: gender (x=1: female, x=0: male)
- lawsuit (1973): $.44 = p(a = 1|x = 0) > p(a = 1|x = 1) = .35$
- 'resolved' by Bickel (1975)[13] (See also Pearl[14])
- u: unobserved department they applied to
- $p(a|x) = \sum_{u=1}^{u=6} p(a|x, u)p(u|x)$
- e.g., gender-blind: $p(a|1) - p(a|0) = p(a|u) \cdot (p(u|1) - p(u|0))$

3.21 confounded approach: quasi-experiments + instruments [15]

- Q: does engagement drive retention? (NYT, NFLX, ...)
  - we don't directly control engagement
  - nonetheless useful since many things can influence it
- Q: does serving in Vietnam war decrease earnings[16]?
  - US didn't directly control serving in Vietnam, either[17]
- requires **strong assumptions**, including linear model

3.22 IV: graphical model assumption

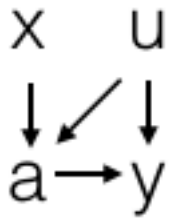


Figure 12: independence assumption

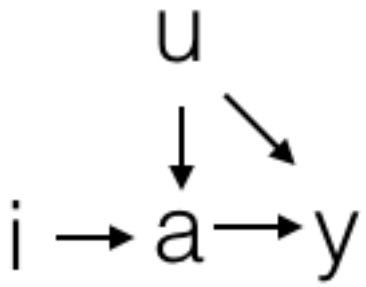


Figure 13: independence assumption

3.23 IV: graphical model assumption (sideways)

3.24 IV: review s/OLS/MOM/ (E is empirical average)

- *a endogenous*
  - e.g., $\exists u$ *s.t.* $p(y|a,x,u), p(a|x,u)$
- linear ansatz: $y = \beta^T a + \epsilon$
- if *a exogenous* (e.g., OLS), use $E[YA_j] = E[\beta^T AA_j] + E[\epsilon A_j]$ (note that $E[A_j A_k]$ gives square matrix; invert for β)
- add *instrument* x uncorrelated with ϵ
- $E[YX_k] = E[\beta^T AX_k] + E[\epsilon]E[X_k]$
- $E[Y] = E[\beta^T A] + E[\epsilon]$ (from ansatz)
- $C(Y, X_k) = \beta^T C(A, X_k)$, not an "inversion" problem, requires "two stage regression"

3.25 IV: binary, binary case (aka "Wald estimator")

- $y = \beta a + \epsilon$
- $E(Y|x) = \beta E(A|x) + E(\epsilon)$, evaluate at $x = \{0, 1\}$
- $\beta = (E(Y|x=1) - E(Y|x=0))/(E(A|x=1) - E(A|x=0))$.

[13] P.J. Bickel, E.A. Hammel and J.W. O'Connell (1975). "Sex Bias in Graduate Admissions: Data From Berkeley". Science 187 (4175): 398–404

[14] Pearl, Judea (December 2013). "Understanding Simpson's paradox". UCLA Cognitive Systems Laboratory, Technical Report R-414.

[15] I thank Sinan Aral, MIT Sloan, for bringing this to my attention

[16] Angrist, Joshua D. "Lifetime earnings and the Vietnam era draft lottery: evidence from social security administrative records." The American Economic Review (1990): 313-336.

[17] cf., George Bush, Donald Trump, Bill Clinton, Dick Cheney...

3.26 bandits: obligatory slide

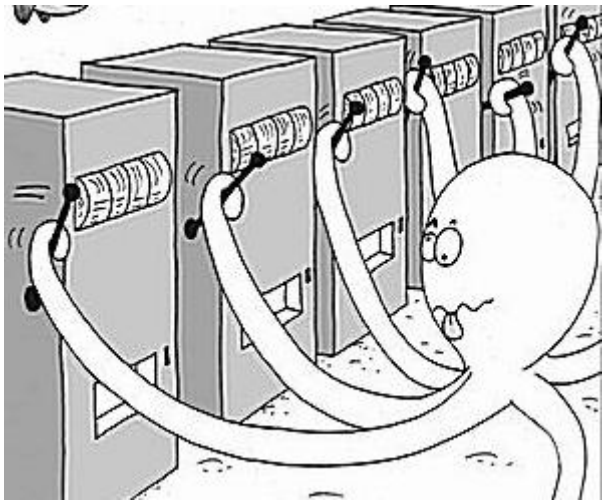


Figure 14: almost all the talks I've gone to on bandits have this image

3.27 bandits

- wide applicability: humane clinical trials, targeting, ...
- replace meetings with code
- requires software engineering to replace decisions with, e.g., Javascript
- most useful if decisions or items get "stale" quickly
- less useful for one-off, major decisions to be "interpreted"

. . .

examples

- ϵ-greedy (no context, aka 'vanilla', aka 'context-free')
- UCB1 (2002) (no context) + LinUCB (with context)
- Thompson Sampling (1933)[18],[19],[20] (general, with or without context)

3.28 TS: connecting w/"generative causal modeling" 0

- WAS $p(y, x, a) = p(y|x, a)p_\alpha(a|x)p(x)$
- These 3 terms were treated by

[18] Thompson, William R. "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples". Biometrika, 25(3–4):285–294, 1933.

[19] AKA "probability matching", "posterior sampling"

[20] cf., "Bayesian Bandit Explorer" (link)

  - response $p(y|a,x)$: avoid regression/inferring using importance sampling
  - policy $p_\alpha(a|x)$: optimize ours, infer theirs
  - (NB: ours was deterministic: $p(a|x) = 1[a = h(x)]$)
  - prior $p(x)$: either avoid by importance sampling or estimate via kernel methods
- In the economics approach we focus on
- $\tau(\ldots) \equiv Q(a = 1, \ldots) - Q(a = 0, \ldots)$ "treatment effect"
- where $Q(a, \ldots) = \sum_y yp(y|\ldots)$

. . .

In Thompson sampling we will generate 1 datum at a time, by

- asserting a parameterized generative model for $p(y|a, x, \theta)$
- using a deterministic but averaged policy

3.29 TS: connecting w/"generative causal modeling" 1

- model true world response function $p(y|a,x)$ parametrically as $p(y|a,x,\theta^*)$
- (i.e., θ^* is the true value of the parameter)[21]
- if you knew θ:
  - could compute $Q(a,x,\theta) \equiv \sum_y yp(y|x,a,\theta^*)$ directly
  - then choose $h(x;\theta) = \text{argmax}_a Q(a,x,\theta)$
  - inducing policy $p(a|x,\theta) = 1[a = h(x;\theta) = \text{argmax}_a Q(a,x,\theta)]$
- idea: use prior data $D = \{y,a,x\}_1^t$ to define *non-deterministic* policy:
  - $p(a|x) = \int d\theta p(a|x,\theta)p(\theta|D)$
  - $p(a|x) = \int d\theta 1[a = \text{argmax}_{a'} Q(a',x,\theta)]p(\theta|D)$
- hold up:
  - Q1: what's $p(\theta|D)$?
  - Q2: how am I going to evaluate this integral?

3.30 TS: connecting w/"generative causal modeling" 2

- Q1: what's $p(\theta|D)$?
- Q2: how am I going to evaluate this integral?
- A1: $p(\theta|D)$ definable by choosing prior $p(\theta|\alpha)$ and likelihood on y given by the (modeled, parameterized) response $p(y|a,x,\theta)$.
  - (now you're not only generative, you're Bayesian.)
  - $p(\theta|D) = p(\theta|\{y\}_1^t, \{a\}_1^t, \{x\}_1^t, \alpha)$
  - $\propto p(\{y\}_1^t|\{a\}_1^t, \{x\}_1^t, \theta)p(\theta|\alpha)$
  - $= p(\theta|\alpha)\Pi_t p(y_t|a_t, x_t, \theta)$

[21] Note that θ is a vector, with components for each action.

  - *warning 1*: sometimes people write "$p(D|\theta)$" but we don't need $p(a|\theta)$ or $p(x|\theta)$ here
  - *warning 2*: don't need historical record of θ_t.
  - (we used Bayes rule, but only in θ and y.)
- A2: evaluate integral by $N = 1$ Monte Carlo
  - take 1 sample "θ_t" of θ from $p(\theta|D)$
  - $a_t = h(x_t; \theta_t) = \text{argmax}_a Q(a, x, \theta_t)$

3.31 That sounds hard.

No, just general. Let's do toy case:

- $y \in \{0, 1\}$,
- no context x,
- Bernoulli (coin flipping), keep track of
  - $S_a \equiv$ number of successes flipping coin a
  - $F_a \equiv$ number of failures flipping coin a

. . .

Then

- $p(\theta|D) \propto p(\theta|\alpha)\Pi_t p(y_t|a_t, \theta)$
- $= \left(\Pi_a \theta_a^{\alpha-1}(1-\theta_a)^{\beta-1}\right)\left(\Pi_{t,a_t}\theta_{a_t}^{y_t}(1-\theta_{a_t})^{1-y_t}\right)$
- $= \Pi_a \theta^{\alpha+S_a-1}(1-\theta_a)^{\beta+F_a-1}$
- $\therefore \theta_a \sim \text{Beta}(\alpha + S_a, \beta + F_a)$

3.32 Thompson sampling: results (2011)

An Empirical Evaluation of Thompson Sampling

Olivier Chapelle
Yahoo! Research
Santa Clara, CA
chap@yahoo-inc.com

Lihong Li
Yahoo! Research
Santa Clara, CA
lihong@yahoo-inc.com

Figure 15: Chaleppe and Li 2011

3.33 TS: words

In the realizable case, the reward is a stochastic function of the action, context and the unknown, true parameter θ^*. Ideally, we would like to choose the action maximizing the expected reward, $\max_a \mathbb{E}(r|a, x, \theta^*)$.

Of course, θ^* is unknown. If we are just interested in maximizing the immediate reward (exploitation), then one should choose the action that maximizes $\mathbb{E}(r|a, x) = \int \mathbb{E}(r|a, x, \theta)P(\theta|D)d\theta$.

But in an exploration / exploitation setting, the probability matching heuristic consists in randomly selecting an action a according to its probability of being optimal. That is, action a is chosen with probability

$$\int \mathbb{I}\left[\mathbb{E}(r|a, x, \theta) = \max_{a'} \mathbb{E}(r|a', x, \theta)\right] P(\theta|D)d\theta,$$

where $\mathbb{I}$ is the indicator function. Note that the integral does not have to be computed explicitly: it suffices to draw a random parameter θ at each round as explained in Algorithm 1. Implementation of the algorithm is thus efficient and straightforward in most applications.

Figure 16: from Chaleppe and Li 2011

3.34 TS: p-code

3.35 TS: Bernoulli bandit p-code[22]

3.36 TS: Bernoulli bandit p-code (results)

3.37 UCB1 (2002), p-code

from Auer, Peter, Nicolo Cesa-Bianchi, and Paul Fischer. "Finite-time analysis of the multiarmed bandit problem." Machine learning 47.2-3 (2002): 235-256.

3.38 TS: with context

3.39 LinUCB: UCB with context

From Li, Lihong, et al. "A contextual-bandit approach to personalized news article recommendation." WWW 2010.

Algorithm 1 Thompson sampling

$D = \emptyset$
for $t = 1, \ldots, T$ **do**
  Receive context x_t
  Draw θ^t according to $P(\theta|D)$
  Select $a_t = \arg\max_a \mathbb{E}_r(r|x_t, a, \theta^t)$
  Observe reward r_t
  $D = D \cup (x_t, a_t, r_t)$
end for

Figure 17: from Chaleppe and Li 2011

Algorithm 2 Thompson sampling for the Bernoulli bandit

Require: α, β prior parameters of a Beta distribution
  $S_i = 0, F_i = 0, \ \forall i$. {Success and failure counters}
  for $t = 1, \ldots, T$ **do**
    for $i = 1, \ldots, K$ **do**
      Draw θ_i according to Beta$(S_i + \alpha, F_i + \beta)$.
    end for
    Draw arm $\hat{\imath} = \arg\max_i \theta_i$ and observe reward r
    if $r = 1$ **then**
      $S_{\hat{\imath}} = S_{\hat{\imath}} + 1$
    else
      $F_{\hat{\imath}} = F_{\hat{\imath}} + 1$
    end if
  end for

Figure 18: from Chaleppe and Li 2011

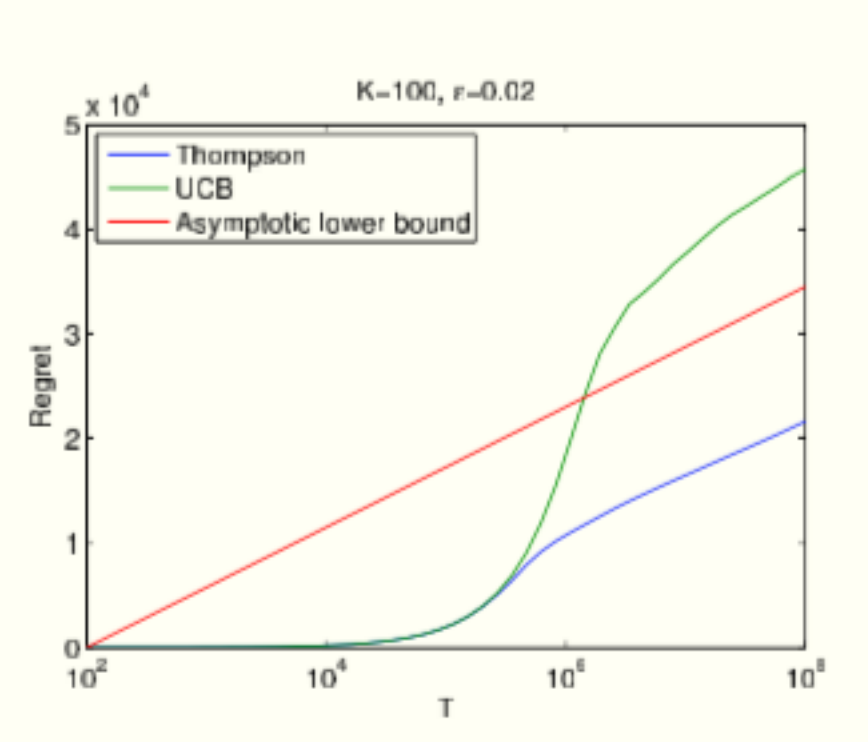


Figure 19: from Chaleppe and Li 2011

Deterministic policy: UCB1.
Initialization: Play each machine once.
Loop:

- Play machine j that maximizes $\bar{x}_j + \sqrt{\frac{2 \ln n}{n_j}}$, where $\bar{x}_j$ is the average reward obtained from machine j, n_j is the number of times machine j has been played so far, and n is the overall number of plays done so far.

Figure 20: UCB1

Algorithm 3 Regularized logistic regression with batch updates

Require: Regularization parameter $\lambda > 0$.

$m_i = 0$, $q_i = \lambda$. {Each weight w_i has an independent prior $\mathcal{N}(m_i, q_i^{-1})$}

for $t = 1, \ldots, T$ **do**

Get a new batch of training data $(\mathbf{x}_j, y_j)$, $j = 1, \ldots, n$.

Find $\mathbf{w}$ as the minimizer of: $\frac{1}{2}\sum_{i=1}^{d} q_i(w_i - m_i)^2 + \sum_{j=1}^{n} \log(1 + \exp(-y_j \mathbf{w}^\top \mathbf{x}_j))$.

$m_i = w_i$

$q_i = q_i + \sum_{j=1}^{n} x_{ij}^2 p_j(1 - p_j)$, $p_j = (1 + \exp(-\mathbf{w}^\top \mathbf{x}_j))^{-1}$ {Laplace approximation}

end for

Figure 21: from Chaleppe and Li 2011

Algorithm 1 LinUCB with disjoint linear models.

Inputs: $\alpha \in \mathbb{R}_+$

for $t = 1, 2, 3, \ldots, T$ **do**

  Observe features of all arms $a \in \mathcal{A}_t$: $\mathbf{x}_{t,a} \in \mathbb{R}^d$

  for all $a \in \mathcal{A}_t$ **do**

    if a is new **then**

      $\mathbf{A}_a \leftarrow \mathbf{I}_d$ (d-dimensional identity matrix)

      $\mathbf{b}_a \leftarrow \mathbf{0}_{d \times 1}$ (d-dimensional zero vector)

    end if

    $\hat{\boldsymbol{\theta}}_a \leftarrow \mathbf{A}_a^{-1} \mathbf{b}_a$

    $p_{t,a} \leftarrow \hat{\boldsymbol{\theta}}_a^\top \mathbf{x}_{t,a} + \alpha \sqrt{\mathbf{x}_{t,a}^\top \mathbf{A}_a^{-1} \mathbf{x}_{t,a}}$

  end for

  Choose arm $a_t = \arg\max_{a \in \mathcal{A}_t} p_{t,a}$ with ties broken arbitrarily, and observe a real-valued payoff r_t

  $\mathbf{A}_{a_t} \leftarrow \mathbf{A}_{a_t} + \mathbf{x}_{t,a_t} \mathbf{x}_{t,a_t}^\top$

  $\mathbf{b}_{a_t} \leftarrow \mathbf{b}_{a_t} + r_t \mathbf{x}_{t,a_t}$

end for

Figure 22: LinUCB

Table 2: CTR regrets on the display advertising data.

Method	TS			LinUCB			ε-greedy			Exploit	Randor
Parameter	0.25	0.5	1	0.5	1	2	0.005	0.01	0.02		
Regret (%)	4.45	3.72	3.81	4.99	4.22	4.14	5.05	4.98	5.22	5.00	31.95

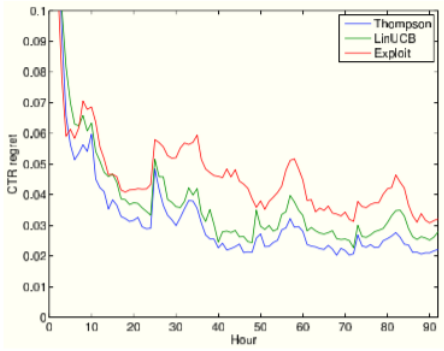


Figure 4: CTR regret over the 4 days test period for 3 algorithms: Thompson sampling with $\alpha = 0.5$, LinUCB with $\alpha = 2$, Exploit-only. The regret in the first hour is large, around 0.3, because the algorithms predict randomly (no initial model provided).

Figure 23: from Chaleppe and Li 2011

THEOREM 2. *Assume that* $I(\theta, \lambda)$ *satisfies* (1.6) *and* (1.7) *and that* Θ *satisfies* (1.9). *Fix* $j \in \{1, \ldots, k\}$, *and define* Θ_j *and* Θ_j^* *by* (2.1). *Let* φ *be any rule such that for every* $\theta \in \Theta_j^*$, *as* $n \to \infty$

$$\sum_{i \neq j} E_\theta T_n(i) = o(n^a) \qquad \text{for every } a > 0, \tag{2.2}$$

where $T_n(i)$, *defined in* (1.2), *is the number of times that the rule* φ *samples from* Π_i *up to stage* n. *Then for every* $\theta \in \Theta_j$ *and every* $\epsilon > 0$,

$$\lim_{n \to \infty} P_\theta\{T_n(j) \geq (1 - \epsilon)(\log n)/I(\theta_j, \theta^*)\} = 1, \tag{2.3}$$

where θ^* *is defined in* (1.4), *and hence*

$$\liminf_{n \to \infty} E_\theta T_n(j)/\log n \geq 1/I(\theta_j, \theta^*).$$

Figure 24: Lai Robbins

3.40 TS: with context (results)

3.41 Bandits: Regret via Lai and Robbins (1985)

3.42 Thompson sampling (1933) and optimality (2013)

Theorem 2. *For any instance* $\Theta = \{\mu_1, ..., \mu_N\}$ *of Bernoulli MAB,*

$$R(T, \Theta) \leq (1 + \epsilon) \sum_{i \neq I^*} \frac{\ln(T)\Delta_i}{KL(\mu_i, \mu^*)} + O(N/\epsilon^2)$$

Recall that we have $\lim_{T\to\infty} \frac{R(T,\Theta)}{\ln(T)} \geq \sum_{i\neq I^*} \frac{\Delta_i}{KL(\mu_i,\mu^*)}$. Above theorem says that Thompson Sampling matches this lower bound. We also have the following problem independent regret bound for this algorithm.

Theorem 3. *For all* Θ,

$$R(T) = \max_{\Theta} R(T, \Theta) \leq O(\sqrt{NT \log T} + N)$$

For proofs of above theorems, refer to [2].

Figure 25: TS result

from S. Agrawal, N. Goyal, "Further optimal regret bounds for Thompson Sampling", AISTATS 2013.; see also Agrawal, Shipra, and Navin Goyal. "Analysis of Thompson Sampling for the Multi-armed Bandit Problem." COLT. 2012 and Emilie Kaufmann, Nathaniel Korda, and R´emi Munos. Thompson sampling: An asymptotically optimal finite-time analysis. In Algorithmic Learning Theory, pages 199–213. Springer, 2012.

3.43 other 'Causalities': structure learning

D. Heckerman. A Tutorial on Learning with Bayesian Networks. Technical Report MSR-TR-95-06, Microsoft Research, March, 1995.

3.44 other 'Causalities': potential outcomes

- model distribution of $p(y_i(1), y_i(0), a_i, x_i)$
- "action" replaced by "observed outcome"
- aka Neyman-Rubin causal model: Neyman ('23); Rubin ('74)
- see Morgan + Winship[23] for connections between frameworks

[22] Note that θ is a vector, with components for each action.

[23] Morgan, Stephen L., and Christopher Winship. *Counterfactuals and causal inference* Cambridge University Press, 2014.

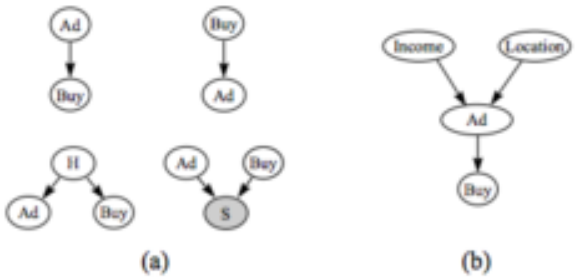


Figure 9: (a) Causal graphs showing for explanations for an observed dependence between *Ad* and *Buy*. The node H corresponds to a hidden common cause of *Ad* and *Buy*. The shaded node S indicates that the case has been included in the database. (b) A Bayesian network for which A causes B is the only causal explanation, given the causal Markov condition.

Figure 26: from heckerman 1995

4 Lecture 4: descriptive modeling @ NYT

4.1 review: (latent) inference and clustering

- what does kmeans mean?
  - given $x_i \in R^D$
  - given $d : R^D \to R^1$
  - assign z_i
- generative modeling gives meaning
  - given $p(x|z, \theta)$
  - maximize $p(x|\theta)$
  - output assignment $p(z|x, \theta)$

4.2 actual math

- define $P \equiv p(x, z|\theta)$
- log-likelihood $L \equiv \log p(x|\theta) = \log \sum_z P = \log E_q P/q$ (cf. importance sampling)
- Jensen's: $L \geq \tilde{L} \equiv E_q \log P/q = E_q \log P + H[q] = -(U - H) = -\mathcal{F}$
  - analogy to free energy in physics
- alternate optimization on θ and on q
  - NB: q step gives $q(z) = p(z|x, \theta)$
  - NB: $\log P$ convenient for independent examples w/ exponential families

- e.g., GMMs: $\mu_k \leftarrow E[x|z]$ and $\sigma_k^2 \leftarrow E[(x-\mu)^2|z]$ are sufficient statistics
- e.g., LDAs: word counts are sufficient statistics

4.3 tangent: more math on GMMs, part 1

Energy U (to be minimized):

- $-U \equiv E_q \log P = \sum_z \sum_i q_i(z) \log P(x_i, z_i) \equiv U_x + U_z$
- $-U_x \equiv \sum_z \sum_i q_i(z) \log p(x_i|z_i)$
- $= \sum_i \sum_z q_i(z) \sum_k 1[z_i = k] \log p(x_i|z_i)$
- define $r_{ik} = \sum_z q_i(z) 1[z_i = k]$
- $-U_x = \sum_i r_{ik} \log p(x_i|k)$.
- Gaussian[24] $\Rightarrow -U_x = \sum_i r_{ik} \left(-\frac{1}{2}(x_i - \mu_k)^2 \lambda_k + \frac{1}{2} \ln \lambda_k - \frac{1}{2} \ln 2\pi\right)$

. . .

simple to minimize for parameters $\vartheta = \{\mu_k, \lambda_k\}$

4.4 tangent: more math on GMMs, part 2

- -$U_x = \sum_i r_{ik} \left(-\frac{1}{2}(x_i - \mu_k)^2 \lambda_k + \frac{1}{2} \ln \lambda_k - \frac{1}{2} \ln 2\pi\right)$
- $\mu_k \leftarrow E[x|k]$ solves $\sum_i r_{ik} = \sum_i r_{ik} x_i$
- $\lambda_k \leftarrow E[(x-\mu)^2|k]$ solves $\sum_i r_{ik} \frac{1}{2}(x_i - \mu_k)^2 = \lambda_k^{-1} \sum_i r_{ik}$

4.5 tangent: Gaussians $\in$ exponential family[25]

- as before, $-U = \sum_i r_{ik} \log p(x_i|k)$
- define $p(x_i|k) = \exp\left(\eta(\theta) \cdot T(x) - A(\theta) + B(x)\right)$
- e.g., Gaussian case [26],
  - $T_1 = x$,
  - $T_2 = x^2$
  - $\eta_1 = \mu/\sigma^2 = \mu\lambda$
  - $\eta_2 = -\frac{1}{2}\lambda = -1/(2\sigma^2)$
  - $A = \lambda\mu^2/2 - \frac{1}{2} \ln \lambda$
  - $\exp(B(x)) = (2\pi)^{-1/2}$
- note that in a mixture model, there are separate η (and thus $A(\eta)$) for each value of z

[24] math is simpler if you work with $\lambda_k \equiv \sigma^{-2}$

[25] NB: Gaussians $\in$ exponential family, GMM $\notin$ exponential family! (Thanks to Eszter Vértes for pointing out this error in earlier title.)

[26] Choosing $\eta(\theta) = \eta$ called 'canonical form'

4.6 tangent: variational joy $\in$ exponential family

- as before, $-U = \sum_i r_{ik} \left(\eta_k^T T(x_i) - A(\eta_k) + B(x_i) \right)$
- $\eta_{k,\alpha}$ solves $\sum_i r_{ik} T_{k,\alpha}(x_i) = \frac{\partial A(\eta_k)}{\partial \eta_{k,\alpha}} \sum_i r_{ik}$ (canonical)
- $\therefore \partial_{\eta_{k,\alpha}} A(\eta_k) \leftarrow E[T_{k,\alpha}|k]$ (canonical)
- nice connection w/physics, esp. mean field theory[27]

4.7 clustering and inference: GMM/k-means case study

- generative model gives meaning and optimization
- large freedom to choose different optimization approaches
  - e.g., hard clustering limit
  - e.g., streaming solutions
  - e.g., stochastic gradient methods

4.8 general framework: E+M/variational

- e.g., GMM+hard clustering gives kmeans
- e.g., some favorite applications:
  - hmm
  - vbmod: arXiv:0709.3512
  - ebfret: ebfret.github.io
  - EDHMM: edhmm.github.io

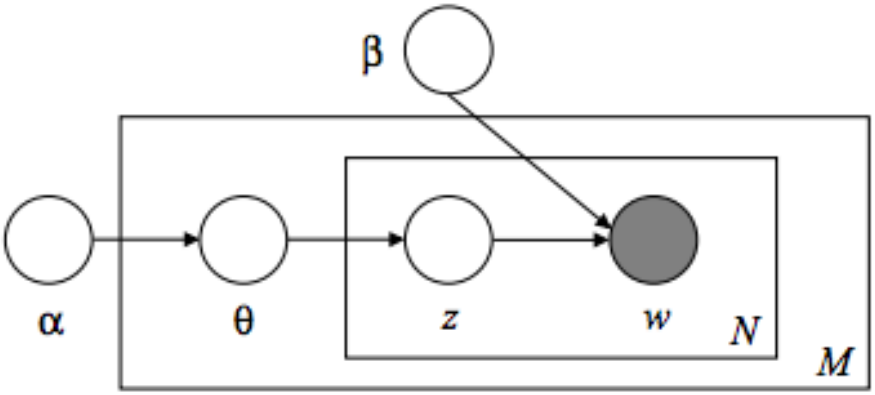


Figure 27: From Blei 2003

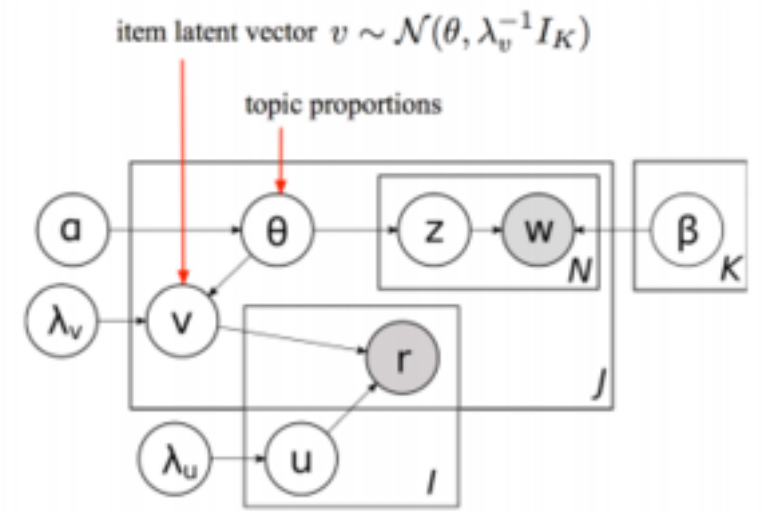


Figure 28: From Blei 2011

$$\min_{U,V} \sum_{i,j} (r_{ij} - u_i^T v_j)^2 + \lambda_u ||u_i||^2 + \lambda_v ||v_j||^2,$$

Figure 29: From Blei 2011

Maximization of the posterior is equivalent to maximizing the complete log likelihood of U, V, $\theta_{1:J}$, and R given λ_u, λ_v and β,

$$\mathcal{L} = -\frac{\lambda_u}{2}\sum_i u_i^T u_i - \frac{\lambda_v}{2}\sum_j (v_j - \theta_j)^T(v_j - \theta_j) \quad (7)$$
$$+ \sum_j \sum_n \log\left(\sum_k \theta_{jk}\beta_{k,w_{jn}}\right) - \sum_{i,j}\frac{c_{ij}}{2}(r_{ij} - u_i^T v_j)^2.$$

Figure 30: From Blei 2011

$$u_i \leftarrow (VC_iV^T + \lambda_u I_K)^{-1}VC_iR_i \quad (8)$$
$$v_j \leftarrow (UC_jU^T + \lambda_v I_K)^{-1}(UC_jR_j + \lambda_v\theta_j). \quad (9)$$

Figure 31: From Blei 2011

$$\begin{aligned}\mathcal{L}(\theta_j) &\geq -\frac{\lambda_v}{2}(v_j - \theta_j)^T(v_j - \theta_j) \\ &\quad + \sum_n \sum_k \phi_{jnk}\left(\log\theta_{jk}\beta_{k,w_{jn}} - \log\phi_{jnk}\right) \\ &= \mathcal{L}(\theta_j, \phi_j). \quad (10)\end{aligned}$$

Figure 32: From Blei 2011

4.9 example application: LDA+topics

4.10 rec engine via CTM [28]

4.11 recall: recommendation via factoring

4.12 CTM: combined loss function

4.13 CTM: updates for factors

4.14 CTM: (via Jensen's, again) bound on loss

5 Lecture 5 data product

5.1 data science and design thinking

- knowing customer
- right tool for right job
- practical matters:
  - munging
  - data ops
  - ML in prod

5.2 Thanks!

Thanks MLSS students for your great questions; please contact me @chrishwiggins or chris.wiggins@{nytimes,gmail}.com with any questions, comments, or suggestions!

[27] read MacKay, David JC. *Information theory, inference and learning algorithms*, Cambridge university press, 2003 to learn more. Actually you should read it regardless.

[28] cf., bit.ly/AlexCTM for NYT blog post on how CTM informs our rec engine