data science @ NYT

Chris Wiggins

Aug 8/9, 2016

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. Thank the organizers!

Lecture 1: overview of ds@NYT

Lecture 2: predictive modeling @ NYT

desc/pred/pres

- caveat: difference between observation and experiment. why?

blossom example

blossom + boosting ('exponential')

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 $\sum_i \mathrm{e}^{-y_i w^T h(x_i)}$

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 $y \in \{-1, +1\}$

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], ...

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

predicting people

- “customer journey” prediction
  - fun covariates
  - observational complication v structural models

predicting people (reminder)

single copy (reminder)

example in CAR (computer assisted reporting)

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

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

[3]By Hiroko Tabuchi, a Pulitzer winner

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

computer assisted reporting

- Impact

Lecture 3: prescriptive modeling @ NYT

the natural abstraction

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

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

maximizing outcome

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

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?

[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.

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

in general

also describes both the ‘exploration’ and ‘exploitation’ distributions

randomized controlled trial

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

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

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 $p_-(y,a,x)$.

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

POISE math: IS+Monte Carlo estimation=ISE

i.e, "importance sampling estimation"

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

POISE math: IS+Monte Carlo estimation=ISE

i.e, "importance sampling estimation"

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

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 $\{y, a, x\}$

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)

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

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: $1[c = d] = 1 - 1[c \neq d]$
- $\therefore E_+(Y) \propto \mathrm{constant} - \sum_i w_i 1[a \neq h(x)]$
- $\therefore$ reduces policy optimization to (weighted) classification

---

[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)

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 $\tilde{L} \equiv \frac{\sum_i y 1[a_i \neq h(x_i)]/p_-(a_i|x_i)}{\sum_i 1[a_i \neq h(x_i)]/p_-(a_i|x_i)}$
- 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

[11] Suggestion by Dan Hsu

POISE punchlines

- allows policy planning even with implicit logged exploration data[12]
- e.g., two hospital story
- "personalized medicine" is also a policy
- abundant data available, under-explored IMHO

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

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)
    - $\tau(x) \equiv E_0(Y|a=1,x) - E_0(Y|a=0,x)$
    - $\equiv Q(a=1,x) - Q(a=0,x)$

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
  $E_p(f) \approx N^{-1} \sum_i \sum_{yax} f(y, a, x)K_1(y|y_i)K_2(a|a_i)K_3(x|x_i)$

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

Helps think about economists' approach:

- $Q(a,x) \equiv E(Y|a,x) = \sum_y yp(y|a,x) = \sum_y y \frac{p_-(y,a,x)}{p_-(a|x)p(x)}$
- $= \frac{1}{p_-(a|x)p(x)} \sum_y yp_-(y,a,x)$
- stratify x using $z(x)$ such that $\cup z = X$, and $\cap z, z' =$
- $n(x) = \sum_i 1[z(x_i) = z(x)]$=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, $K_1(y|y_i) = \delta_{y,y_i}$, $K_2(a|a_i) = \delta_{a,a_i}$

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

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

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,. . .

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

causality, as understood in marketing

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

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

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: $p(a|1) - p(a|0) = p(a|u) \cdot (p(u|1) - p(u|0))$

[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.

confounded approach: quasi-experiments + instruments [17]

- 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[15]?
  - US didn't directly control serving in Vietnam, either[16]
- requires **strong assumptions**, including linear model

---

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

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

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

IV: graphical model assumption

IV: graphical model assumption (sideways)

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 $E[YA_j] = E[\beta^T A A_j] + E[\epsilon A_j]$ (note that $E[A_j A_k]$ gives square matrix; invert for β)
- add *instrument* x uncorrelated with ϵ
- $E[YX_k] = E[\beta^T A X_k] + E[\epsilon]E[X_k]$
- $E[Y] = E[\beta^T A] + E[\epsilon]$ (from ansatz)
- $C(Y, X_k) = \beta^T C(A, X_k)$, not an "inversion" problem, requires "two stage regression"

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

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

bandits: obligatory slide

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"

[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)

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)

---

[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)

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
  - 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 $Q(a, \ldots) = \sum_y yp(y|\ldots)$

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
  - 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

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 $Q(a, x, \theta) \equiv \sum_y y p(y|x, a, \theta^*)$ directly
  - then choose $h(x; \theta) = \mathrm{argmax}_a Q(a, x, \theta)$
  - inducing policy $p(a|x, \theta) = 1[a = h(x; \theta) = \mathrm{argmax}_a Q(a, x, \theta)]$
- idea: use prior data $D = \{y, a, x\}_1^t$ to define *non-deterministic* policy:
  - $p(a|x) = \int d\theta p(a|x, \theta) p(\theta|D)$
  - $p(a|x) = \int d\theta 1[a = \mathrm{argmax}_{a'} Q(a', x, \theta)] p(\theta|D)$
- hold up:
  - Q1: what's $p(\theta|D)$?
  - Q2: how am I going to evaluate this integral?

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

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.)
  - $p(\theta|D) = p(\theta|\{y\}_1^t, \{a\}_1^t, \{x\}_1^t, \alpha)$
  - $\propto p(\{y\}_1^t|\{a\}_1^t, \{x\}_1^t, \theta)p(\theta|\alpha)$
  - $= p(\theta|\alpha)\Pi_t p(y_t|a_t, x_t, \theta)$
  - *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 $p(\theta|D)$
  - $a_t = h(x_t; \theta_t) = \mathrm{argmax}_a Q(a, x, \theta_t)$

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

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

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

Thompson sampling: results (2011)

TS: words

TS: p-code

TS: Bernoulli bandit p-code[22]

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

TS: Bernoulli bandit p-code (results)

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.

TS: with context

LinUCB: UCB with context

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

TS: with context (results)

Bandits: Regret via Lai and Robbins (1985)

Thompson sampling (1933) and optimality (2013)

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.

other 'Causalities': structure learning

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

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

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

Lecture 4: descriptive modeling @ NYT

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 $p(z|x,\theta)$

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

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]
  $\Rightarrow -U_x = \sum_i r_{ik} \left(-\frac{1}{2}(x_i - \mu_k)^2 \lambda_k + \frac{1}{2} \ln \lambda_k - \frac{1}{2} \ln 2\pi \right)$

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

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 $\vartheta = \{\mu_k, \lambda_k\}$

---

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

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 $\sum_i r_{ik} \frac{1}{2}(x_i - \mu_k)^2 = \lambda_k^{-1} \sum_i r_{ik}$

tangent: Gaussians $\in$ exponential family[26]

- 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 [25],
  - $T_1 = x$,
  - $T_2 = x^2$
  - $\eta_1 = \mu/\sigma^2 = \mu\lambda$
  - $\eta_2 = -\frac{1}{2}\lambda = -1/(2\sigma^2)$
  - $A = \lambda\mu^2/2 - \frac{1}{2}\ln\lambda$
  - $\exp(B(x)) = (2\pi)^{-1/2}$
- note that in a mixture model, there are separate η (and thus $A(\eta)$) for each value of z

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

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

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]

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

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

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

example application: LDA+topics

rec engine via CTM [28]

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

recall: recommendation via factoring

CTM: combined loss function

CTM: updates for factors

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

Lecture 5 data product

data science and design thinking

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

Thanks!

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