10/28/14

To clear up:

- lateness to class?
- late on assignments?
- ~~not showing~~ missed class? → tell me what's going on.
- missed assignment? { min. completion / contract for a week. (LL - need to assign grades)

Today: question genre: "what makes two types of lang. different?"
altruistic vs. not, funny v. not, in ~~one ti~~ before an event vs. after an event, dem. vs. repub.

- many approaches; today, an example (note) ~~Bayesian~~ statistical approach from a Bayesian perspective (as opposed to a classification perspective)
  → goal: inspire you to develop new tests.

Desiderata:

refinement: we want significant differences: → ... where diffs (probably) not due to chance?

(explained: avg |V| = 60K ⇒ # possible 5-grams = $(60K)^5$; ~~that~~ not enough docs to contain them all)

⇒ need a way to talk about things being due to chance

- ~~will want to be able to estimate variance of an estimat~~

(a) • want estimates of the variance of our 'difference' statistics.

(b) in lang, generally never enough data. (~~related~~ this is a related point) and this will certainly be the case for your settings, where you're generally looking @ sthg "tricky"/restricted

(c) want to preserve, b/c of our new domains, not pre-filter features, as far as computationally possible — how do you know what will be distinguishing

< ex: um/uh for gender, "still" for sentiment, ? → keep punctuation, 2/4 aren't #'s in Twitter (FMs!) >

< is Zipf's law an example of this? >

Zipf's law (gloss): ~~sm~~ rare ~~events~~ words ~~make up~~ comprise surprisingly large fraction of data
→ prob of x prop. to $\frac{1}{(\text{rank of } x)^{\gamma}}$. So tail is "heavier" than an exp. decr-ease

idea: use priors ~~prob~~ to express general knowledge about lang, [even if never seen a dem or repub doc, you know "the" is gonna be relatively frequent] ameliorate the sparse data problem.

< foreshadowing: "we" know an estimate of the variance of the ~~posterior~~ log-odds ratio (in particular for the ~~MLE~~ MLE multinomial given a (posterior) Dirichlet distribution on the multinomial parameters) for a ~~the MLE~~ multinomial distribution, ~~ie~~ w/ Dirichlet prior >

⇒ that's where we're gonna be headed
main idea of Monroe/~~Colaresi/Quinn~~ et al → distinguished political scientists who've also collaborated w/ computer scientists...
"Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict", Political Analysis 2008

---

contrasts: statistical { ~~reg~~ logistic regression coefficients (e.g., Mishra & Gilbert '14) → altho' these turn out to be log-odds ratios = expected change in log-odds for a one-unit change in corresponding standardized var, other vars held constant! (if no interaction terms)
stat-based coin-flipping for burstiness detection (Kleinberg '02)
→ underlying Markov model
~~[statistical but no explicit prior or estimate of variance~~.

ex: ref: FAQ: how do I interpret odds ratios in logistic regression?

[(c) in "real" research: use held-out set for exploring diffs, if want to be proper and plan to do classification experiments]

(True, it's sad that you lose even more data this way)

Setting: ~~Two sequences~~ ~~data = lang seq 1: $\vec{s} = s_1, \ldots s_i$~~

two samples, one from one language and the other from the other, say Dem. Repub. on topic of abortion

~~$\vec{s} = s_1 s_2 \cdots s_i \cdots s_n$~~ : ~~seq s_i~~ $v_i \in V$ the "vocabulary"

~~$S' = s_1' s_2' \cdots s_{n'}'$~~

[can generalize to features, like POS tags, but want some notion of plausible independence]

running example, go?

like to be able to see that, say, a word like "pro-life" has diff. prob of occ. w.r.t. the two ~~under~~ generation processes.

Later

$\boxed{\text{Assume words } \cancel{s_i's} \text{ ~~are~~ drawn ~~indep~~, i.i.d. from a multinomial } \cancel{\theta, \theta'} \text{ unk. w. params } \vec{\theta}^D = (\theta_1^D, \ldots, \theta_{|V|}^D),\ \theta^R = (\theta_1^R, \ldots, \theta_{|V|}^R)}$

consider the count histograms: – our summary of the observed evidence:

$(c_1^D, c_2^D, \ldots, c_{|V|}^D)$ vs. $(c_1^R, c_2^R, \ldots, c_{|V|}^R)$.

<assumed to be a stream – concatenate the docs>

<following> MCQ: 1st gather intuitions and establish notation by looking @ single statistics.

[3.2.1: diff. in frequencies:] rank v_i by $c_i^D - c_i^R$? → give away.

- what if one side speaks more? ✗
- be carefully about cherry-picking (pg. 376)
  - ranking given in paper makes ~~it~~ more clear ~~the~~ : "the", "of", "so"

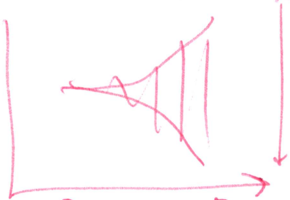
    probs

3.2.2 diff. in proportions: normalize: $\dfrac{c_i^D}{\sum_j c_j^D} - \dfrac{c_i^R}{\sum_j c_j^R}$

"rf_i^D (rel. freq.)" "rf^R (rel. freq)".

See Fig 1, pg 377

great visualization idea by MCQ! →



ranking of top 20 D, top 20 R

Frequency

~~⇒ most freq words are getting most weight~~

most weighted words are all high-freq.

claim: variation in high-freq words not accounted for:

low freq diffs: $\frac{2}{N_D} - \frac{1}{N} = \frac{1}{N}$; high-freq diffs: $\frac{2000}{N} - \frac{1990}{N} = \frac{10}{N} \gg \frac{1}{N}$

<note restricted range ~~actually, not~~ } not needed, but build intuition (or reinforce)

b/c we know ~~we~~ we're gonna want to get to a log-odds ratio:
<foreshadow: research drama — this won't go well, but is leading to sthg good >

let's consider odds

$$\frac{rf_i^D}{1-rf_i^D}$$

~~scale~~ scale is ~~...~~ outside (0,1) - more extreme gives bigger #.
<consider "4:1" odds that the word will appear for Dem".

~~log-odds-ratio~~

log-odds-ratio: $\log\left(\frac{rf_i^D/(1-rf_i^D)}{rf_i^R/(1-rf_i^R)}\right)$

- if equal rel freqs, get 0.
- informative sign (why are Dem's ⊕ ⌣ ?)
- problem for unseen words.

could just smooth by adding in small fake counts to everything, normalize appropriately.

see fig 2 - high-wt words are all low-freq, b/c ratios are more extreme for low counts (same 'noodling' 1/N example as before)

~~let's go back to model-based approach:~~
~~→ use notation from previous page that was deferred.~~

- Should we keep trying ad hoc fixes? Or try sthg more "principled"?
let's try to explicitly take variance into acct: if we see a big log-odds ratio, is that just by chance?
need: what is the variance of the log-odds ratio?
→ what is its distribution?

} why? b/c then can do a z-score test: Remember 1.96??

model: multinomial generation (use notation that was deferred from previous page)
- natural, ~~altho~~ for independent rolls of dies or draws of words.
(altho not true ('Noriegas', Ken Church)

⋮

what do we know about $\vec{\theta}^D, \vec{\theta}^R$.

express prior beliefs: Dirichlet is very convenient: conjugate prior for multinomial.
(not that we know any other priors on multinomials).

params $\alpha_1, \alpha_2, \ldots, \alpha_{|V|}$; call $\alpha_0 = \sum \alpha_i$, require $\alpha_i > 0$.
(maybe don't need to introduce

~~prefers multinomials where the θ_i's are roughly normalized~~

probably better to have just talked about the mean.

mode: $\theta_i = \frac{\alpha_i - 1}{\sum_j (\alpha_j - 1)}$ - so, likes multinomials like this, → if $\alpha_i > 1$ (Albert & Denis, 2012 as one 'random' reference).

mult. w/ params

(Note: the "pts" being drawn here represent multinomials, "an urn of urns", Matthew said!

variance ↓ as α_i's increase.

described as the "normalization" trick.

mean: $\theta_i = \frac{\alpha_i}{\alpha_0}$

"concentration" rel diffs specify expected value.

already guessed use google n-gram corpus

showed: fig 4 - uninformative prior
fig 5 - informative prior.
(~~very~~ definitely made an impression)

posterior $P(\vec{\theta}|\vec{c})$ prop to $P(\vec{c}|\vec{\theta})P(\vec{\theta})$ → is Dir w/ $\alpha_i' = \alpha_i + c_i$!