Whence Linguistic Data?

Bob Carpenter

Alias-i, Inc.

From the Armchair ...

A (computational) linguist in 1984

... to the Observatory

A (computational) linguist in 2010

Supervised Machine Learning

1. Define coding standard mapping inputs to outputs, e.g.:
   - English word $\rightarrow$ stem
   - newswire text $\rightarrow$ person name spans
   - biomedical text $\rightarrow$ genes mentioned
2. Collect inputs and code “gold standard” training data
3. Develop and train statistical model using data
4. Apply to unseen inputs

Coding Bottleneck

- Bottleneck is collecting training corpus
- Commericial data's expensive (e.g. LDA, ELRA)
- Academic corpora typically restrictively licensed
- Limited to existing corpora
- For new problems, use: self, grad students, temps, interns, ...

- Crowdsourcing to the rescue (e.g. Mechanical Turk)

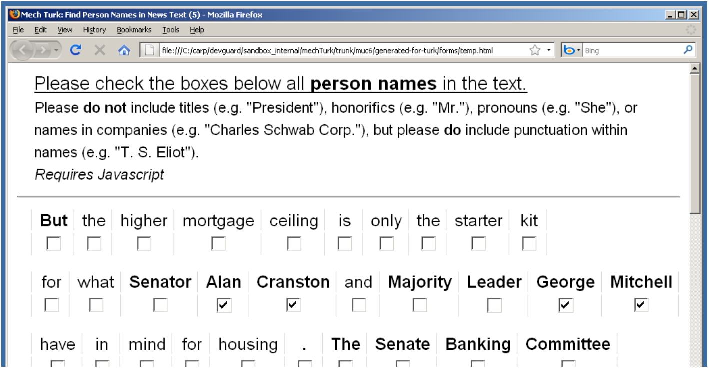
Case Studies

(Mechanical Turked, but same for “experts”.)

Amazon's Mechanical Turk (and its Like)

- "Crowdsourcing" Data Collection
- Provide web forms (or applets) to users
- Users choose tasks to complete
- We can give them a qualifying/training test
- They fill out a form per task and submit
- We pay them through Amazon
- We get the results in a CSV spreadsheet

Case 1: Named Entities



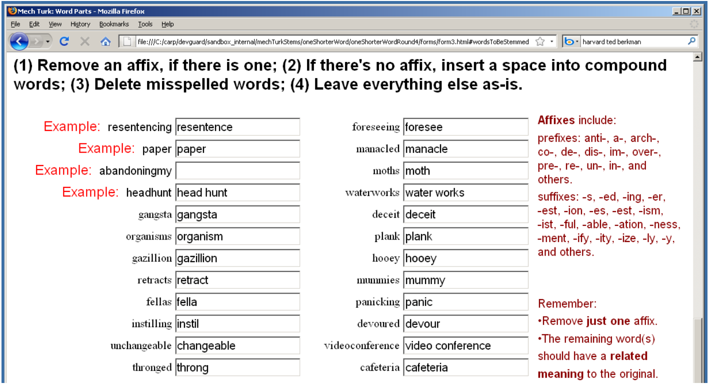
Named Entities Worked

- Conveying the coding standard
  - official MUC-6 standard dozens of pages
  - examples are key
  - (maybe a qualifying exam)
- User Interface Problem
  - highlighting with mouse too fiddly (see Fitts' Law)
  - one entity type at a time (vs. pulldown menus)
  - checkboxes (vs. highlighting spans)

Discussion: Named Entities

- 190K tokens, 64K capitalized, 4K names
- 10 annotators per token
- 100+ annotators, varying numbers of annotations
- Less than a week at 2 cents/400 tokens (US$95)
- Turkers overall better than LDC data
  - Correctly Rejected: `Webster's, Seagram, Du Pont, Buick-Cadillac, Moon, erstwhile Phineas Foggs`
  - Incorrectly Accepted: `Tass`
  - Missed Punctuation: `J E. ‘‘Buster’’ Brown`
- Many Turkers no better than chance

Case 2: Morphological Stemming



Mech Turk: Word Parts - Mozilla Firefox

File Edit View History Bookmarks Tools Help

file:///C:/carp/devguard/sandbox_internal/mechTurkStems/oneShorterWord/oneShorterWordRound4/forms/form3.html#wordsToBeStemmed

harvard ted berkman

(1) Remove an affix, if there is one; (2) If there's no affix, insert a space into compound words; (3) Delete misspelled words; (4) Leave everything else as-is.

Example: resentencing	resentence	foreseeing	foresee
Example: paper	paper	manacled	manacle
Example: abandoningmy		moths	moth
Example: headhunt	head hunt	waterworks	water works
gangsta	gangsta	deceit	deceit
organisms	organism	plank	plank
gazillion	gazillion	hooey	hooey
retracts	retract	mummies	mummy
fellas	fella	panicking	panic
instilling	instil	devoured	devour
unchangeable	changeable	videoconference	video conference
thronged	throng	cafeteria	cafeteria

Affixes include:

prefixes: anti-, a-, arch-, co-, de-, dis-, im-, over-, pre-, re-, un-, in-, and others.

suffixes: -s, -ed, -ing, -er, -est, -ion, -es, -est, -ism, -ist, -ful, -able, -ation, -ness, -ment, -ify, -ity, -ize, -ly, -y, and others.

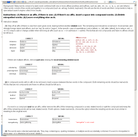
Remember:

•Remove **just one** affix.

•The remaining word(s) should have a **related meaning** to the original.

Morphological Stemming Worked

- Three iterations on coding standard
  - simplified task to one stem
- Four iterations on final standard instructions
  - added previously confusing examples



- Added qualifying test

Case 3: Gene Linkage

Here is your article:

Biochemical, phenotypic and neurophysiological characterization of a genetic mouse model of RSH/Smith--Lemli--Opitz syndrome.

The RSH/Smith--Lemli--Opitz syndrome (RSH/SLOS) is a human autosomal recessive syndrome characterized by multiple malformations, a distinct behavioral phenotype with autistic features and mental retardation. RSH/SLOS is due to an inborn error of cholesterol biosynthesis caused by mutation of the 3 beta-hydroxysterol Delta(7)-reductase gene. To further our understanding of the developmental and neurological processes that underlie the pathophysiology of this disorder, we have developed a mouse model of RSH/SLOS by disruption of the 3 beta-hydroxysterol Delta(7)-reductase gene. Here we provide the biochemical, phenotypic and neurophysiological characterization of this genetic mouse model. As in human patients, the RSH/SLOS mouse has a marked reduction of serum and tissue cholesterol levels and a marked increase of serum and tissue 7-dehydrocholesterol levels. Phenotypic similarities between this mouse model and the human syndrome include intra-uterine growth retardation, variable craniofacial anomalies including cleft palate, poor feeding with an uncoordinated suck, hypotonia and decreased movement. Neurophysiological studies showed that although the response of frontal cortex neurons to the neurotransmitter gamma-amino-n-butyric acid was normal, the response of these same neurons to glutamate was significantly impaired. This finding provides insight into potential mechanisms underlying the neurological dysfunction seen in this human mental retardation syndrome and suggests that this mouse model will allow the testing of potential therapeutic interventions.

Example: Genes: 1024, 5568, 701

Genes: Type your Gene ID's he

(If there are none, leave as-is.)

Remember, the text highlightin genes may or **may not** be usef text!!

Suggested Genes: (Official Name | nickname | EntrezGene ID)

Human Genes:	Mouse Genes:	Rat Genes:
	Plp1: | rsh | 18823	**Plp**: | 24943
DHCR7: | SLOS | 1717	**Dhcr7**: | 13360	**Dhcr7**: | 64191
PSMB6: | DELTA | 5694	**Psmb6**: | 19175	**Psmb6**: | 29666
YY1: | DELTA | 7528	**Yy1**: | 22632	**Yy1**: | 24919
DLL1: | Delta | 28514	**Dll1**: | 13388	**Dll1**: | 84010

Please give us feedback: This is a test run!

- How certain are you of your answers?
- Do you understand what EntrezGene is?
- Were there any phrases that looked like genes but y ones?
- Were there any gene mentions where you could not discussed?
- Did the instructions make sense? Do you have any
- Other comments/suggestions?

Type feedback here.

Gene Linkage Failed

- **Could** get Turkers to pass qualifier
- **Could not** get Turkers to take task even at $1/hit
- Doing coding ourselves (5-10 minutes/HIT)

- How to get Turkers do these complex tasks?
  - Low concentration tasks done quickly
  - Compatible with studies of why Turkers Turk

κ Statistics

κ is "Chance-Adjusted Agreement"

$$\kappa(A, E) = \frac{A - E}{1 - E}$$

- A is agreeement rate
- E is chance agreement rate

- Industry standard
- Attempts to adjust for difficulty of task
- κ above arbitrary threshold considered "good"

Problems with $\mathcal{K}$

- κ intrinsically a pairwise measure
- κ only works for subset of shared annotations
- Not used in inference after calculation
  - κ doesn't predict corpus accuracy
  - κ doesn't predict annotator accuracy
- κ reduces to agreement for hard problems
  - $\lim_{E \to 0} \kappa(A, E) = A$

Problems with $\mathcal{K}$ (cont)

- κ assumes annotators all have same accuracies
- κ assumes annotators are unbiased
  - if biased in same way, κ too high
- κ assumes 0/1 items same value
  - common: low prevalence, high negative agreement
- κ typically estimated without variance component
- κ assumes annotations for an item are uncorrelated
  - items have correlated errors, κ too high

Inferring Gold Standards

Voted Gold Standard

- Turkers vote
- Label with majority category
- Censor if no majority

- This is also NLP standard
- Sometimes adjudicated
  - no reason to trust result

Some Labeled Data

- Seed the data with cases with known labels
- Use known cases to estimate coder accuracy
- Vote with adjustment for accuracy
- Requires relatively large amount of items for
  - estimating accuracies well
  - liveness for new items
- Gold may not be as pure as requesters think
- Some preference tasks have no “right” answer
  - e.g. Dolores Labs’: Bing vs. Google, Facestat, Colors, ...

Estimate Everything

- Gold standard labels
- Coder accuracies
  - sensitivity = TP/(TP+FN) (false negative rate; misses)
  - specificity = TN/(TN+FP) (false positive rate; false alarms)
    - unlke precision, but like κ, uses TN information
  - imbalance indicates bias; high values accuracy
- Coding standard difficulty
  - average accuracies
  - variation among coders
- Item difficulty (important; needs many annotations)

Benefits of (Bayesian) Estimation

- More accurate than voting with threshold
  - largest benefit with few Turkers/item
  - evaluated with known "gold standard"
- May include gold standard cases (semi-supervised)
- Full Bayesian posterior inference
  - probabilistic "gold standard"
  - compatible with probabilistic learning, esp. Bayesian
  - use uncertainty for (overdispersed) downstream inference

Why Task Difficulty for Smoothing?

- What's your estimate for:
  - a baseball player who goes 5 for 20? or 50 for 200?
  - a market that goes down 9 out of 10 days?
  - a coin that lands heads 3 out of 10 times?
  - ...
  - an annotator who's correct for 10 of 10 items?
  - an annotator who's correct in 171 of 219 items?
  - . . .
- Hierarchical model inference for accuracy prior
  - Smooths estimates for coders with few items
  - Supports (multiple) comparisons of accuracies

Is a 24 Karat Gold Standard Possible?

- Or is it fool's gold?
- Some items are marginal given coding standard
  - 'erstwhile Phineas Phoggs' (person?)
  - 'the Moon' (location?)
  - stem of 'butcher' ('butch'?)
- Some items are underspecified in text
  - 'New York' (org or loc?)
  - 'fragile X' (gene or disease?)
  - 'p53' (gene vs. protein vs. family, which species?)
  - operon or siRNA transcribed region (gene or ?)

Traditional Approach to Disagreeement

- Traditional approaches either
  - censor disagreements, or
  - adjudicate disagreements (revise standard).
- Adjudication may not converge
- But, posterior uncertainty can be modeled

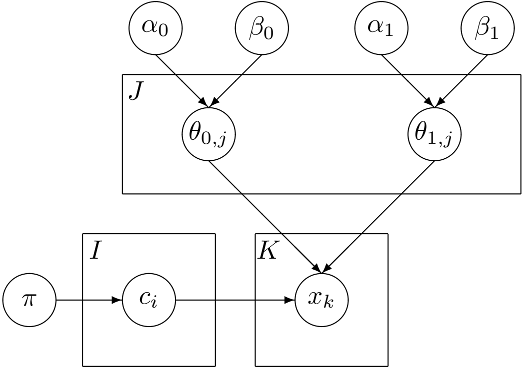
Statistical Inference Model

Strawman Binomial Model

- Prevalence π : chance of "positive" outcome
- $\theta_{1,j}$: annotator j's sensitivity = TP/(TP+FN)
- $\theta_{0,j}$: annotator j's specificity = TN/(TN+FP)
- Sensitivities, specifities same ($\theta_{1,j} = \theta_{0,j'}$)
- Maximum likelihood estimation (or hierarchical prior)

- Hypothesis easily rejected by by χ^2
  - look at marginals (e.g. number of all-1 or all-0 annotations)
  - overdispersed relative to simple model

Beta-Binomial "Random Effects"



Sampling Notation

Label x_k by annotator i_k for item j_k

$$
\begin{aligned}
\pi &\sim \mathsf{Beta}(1,1) \\
c_i &\sim \mathsf{Bernoulli}(\pi) \\
\theta_{0,j} &\sim \mathsf{Beta}(\alpha_0,\beta_0) \\
\theta_{1,j} &\sim \mathsf{Beta}(\alpha_1,\beta_1) \\
x_k &\sim \mathsf{Bernoulli}(c_{i_k}\theta_{1,j_k} + (1-c_{i_k})(1-\theta_{0,j_k}))
\end{aligned}
$$

- $\mathsf{Beta}(1,1) = \mathsf{Uniform}(0,1)$
- Maximum Likelihood: $\alpha_0 = \alpha_1 = \beta_0 = \beta_1 = 1$

Hierarchical Component

- Estimate accuracy priors (α, β)
- With diffuse hyperpriors:

$$\begin{aligned}
\alpha_0/(\alpha_0+\beta_0) &\sim \mathsf{Beta}(1,1) \\
\alpha_0+\beta_0 &\sim \mathsf{Pareto}(1.5) \\
\alpha_1/(\alpha_1+\beta_1) &\sim \mathsf{Beta}(1,1) \\
\alpha_1+\beta_1 &\sim \mathsf{Pareto}(1.5)
\end{aligned}$$

$note$: $\mathsf{Pareto}(x|1.5) \propto x^{-2.5}$

- Infers appropriate smoothing
- Estimates annotator population parameters

Gibbs Sampling

- Estimates full posterior distribution
  - Not just variance, but shape
  - Includes dependencies (covariance)
- Samples $\theta^{(n)}$ support plug-in predictive inference

$$p(y'|y) = \int p(y'|\theta)\; p(\theta|y)\; d\theta \approx \frac{1}{N} \sum_{n<N} p(y'|\theta^{(n)})$$

- Robust (compared to EM)
- Requires sampler for conditionals (automated in BUGS)

BUGS Code

```
model {
  pi ~ dbeta(1,1)
  for (i in 1:I) {
    c[i] ~ dbern(pi)
  }
  for (j in 1:J) {
    theta.0[j] ~ dbeta(alpha.0,beta.0) I(.4,.99)
    theta.1[j] ~ dbeta(alpha.1,beta.1) I(.4,.99)
  }
  for (k in 1:K) {
    bern[k] <- c[ii[k]] * theta.1[jj[k]]
             + (1 - c[ii[k]]) * (1 - theta.0[jj[k]])
    xx[k] ~ dbern(bern[k])
  }
  acc.0 ~ dbeta(1,1)
  scale.0 ~ dpar(1.5,1)  I(1,100)
  alpha.0 <- acc.0 * scale.0
  beta.0 <- (1-acc.0) * scale.0
  acc.1 ~ dbeta(1,1)
  scale.1 ~ dpar(1.5,1) I(1,100)
  alpha.1 <- acc.1 * scale.1;
  beta.1 <- (1-acc.1) * scale.1
}
```

Calling BUGS from R

```
library("R2WinBUGS")

data <- list("I","J","K","xx","ii","jj")

parameters <- c("c", "pi","theta.0","theta.1",
                "alpha.0", "beta.0", "acc.0", "scale.0",
                "alpha.1", "beta.1", "acc.1", "scale.1")

inits <- function() {
  list(pi=runif(1,0.7,0.8),
       c=rbinom(I,1,0.5),
       acc.0 <- runif(1,0.9,0.9),
       scale.0 <- runif(1,5,5),
       acc.1 <- runif(1,0.9,0.9),
       scale.1 <- runif(1,5,5),
       theta.0=runif(J,0.9,0.9),
       theta.1=runif(J,0.9,0.9))  }

anno <- bugs(data, inits, parameters,
             "c:/carp/devguard/sandbox/hierAnno/trunk/R/bugs/beta-binomial-anno.bug",
              n.chains=3, n.iter=500, n.thin=5,
              bugs.directory="c:\\WinBUGS\\WinBUGS14")
```

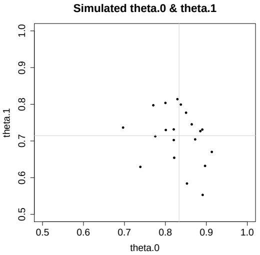
Simulated Data

Simulation Study

- Simulate data (with reasonable model settings)
- Test sampler's ability to fit
- Parameters
  - 20 annotators, 1000 items
  - 50% missing annotations at random
  - prevalence $\pi = 0.2$
  - specificity prior $(\alpha_0, \beta_0) = (40, 8)$ (83% accurate, medium var)
  - sensitivity prior $(\alpha_1, \beta_1) = (20, 8)$ (72% accurate, high var)

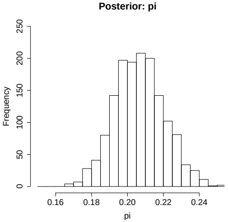
Simulated Sensitivities / Specificities

- Crosshairs at prior mean
- Realistic simulation compared to (estimated) real data

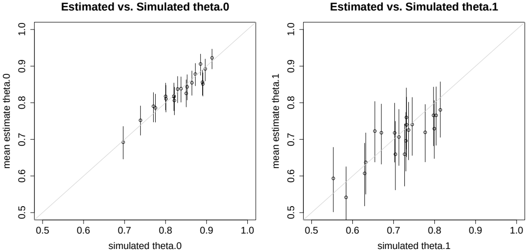


Prevalence Estimate

- Simulated with $\pi = 0.2$
  - sample mean c_i was 0.21
- Estimand of interest in epidemiology (or sentiment)



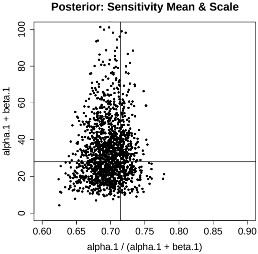
Sensitivity / Specificity Estimates

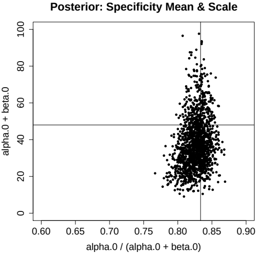


- Posterior mean and 95% intervals
- Diagonal is perfect estimation
- More uncertainty for sensitivity (more data w. $\pi = 0.2$)

Sens / Spec Hyperprior Estimates

Posterior samples $\alpha^{(n)}, \beta^{(n)}$; cross-hairs at known vals.





- Note skew to high scale (low variance)
- Estimates match sampled means

Real Data

5 Dentists Diagnosing Caries

Dentists	*Count*	*Dentists*	*Count*	*Dentists*	*Count*
00000	1880	10000	22	00001	789
10001	26	00010	43	10010	6
00011	75	10011	14	00100	23
10100	1	00101	63	10101	20
00110	8	10110	2	00111	22
10111	17	01000	188	11000	2
01001	191	11001	20	01010	17
11010	6	01011	67	11011	27
01100	15	11100	3	01101	85
11101	72	01110	8	11110	1
01111	56	11111	100		

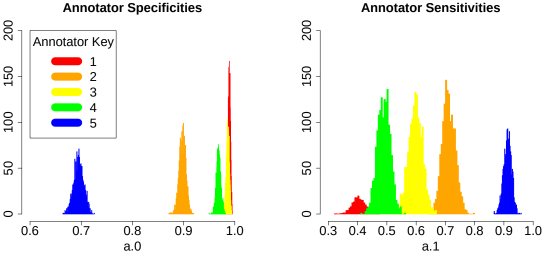
Estimands of Interest

- π: Prevalence of caries
- c_i: 1 if patient i has caries; 0 otherwise
- $\theta_{1,j}$: Sensitivity of dentist j [TP/(TP+FN)]
- $\theta_{0,j}$: Specificity of dentist j [TN/(TN+FP)]
  - can compute precision [TP/(TP+FP)]
  - precision + recall (sensitivity) not complete [no FN]

- task difficulty — priors on θ predict new annotators
- item difficulty

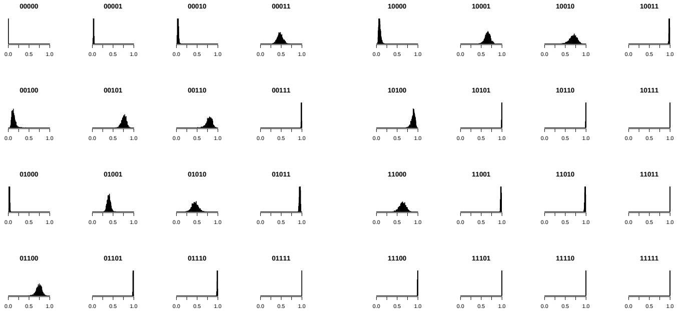
Posteriors for Dentist Accuracies

- In beta-binomial by annotator model



- Posterior density vs. point estimates (e.g. mean)

Posteriors for Dentistry Data Items



Accounts for bias, so very different from simple vote!

Marginal Evaluation

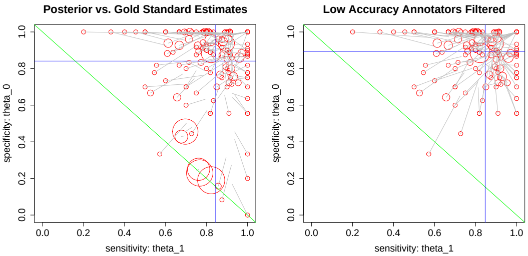
- Common eval in epidemiology
- Models without sensitivity/specificity by annotator underdispersed

Positive Tests	*Frequency*	*Posterior Quantiles* .025	.5	.975
0	1880	1818	1877	1935
1	1065	1029	1068	1117
2	404	385	408	434
3	247	206	227	248
4	173	175	193	212
5	100	80	93	109

Textual Entailment Data

- Collected by Snow et al. using Mechnical Turk
- Recreates a popular linguistic data set (Dagan et al.'s RTE-1)
- $Text$: Microsoft was established in Italy in 1985.
  $Hypothesis$: Microsoft was established in 1985.
- Binary responses true/false
- "Gold Standard" was pretty bad

Estimated vs. “Gold” Accuracies

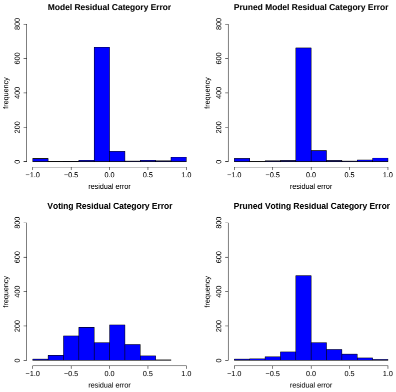


- Diagonal green at chance (below is adversarial)
- blue lines at estimated prior means
- Circle area is items annotated, center at “gold standard” accuracy, lines to estimated accuracy (note pull to prior)

Annotator Pool Estimates

- Gold-standard balanced (50% prevalence)
- Posterior 95
  - Prevalence (.45,.52)
  - Specificity (.81,.87)
  - Sensitivity (.82,.87)
- Posterior sensitivity 95%
  - 39% of annotators no better than chance
  - more than 50% of annotations from spammers
  - has little effect on inference

Residual Category Errors



- Many residual errors in gold standard, not Turkers

Modeling Item Difficulty

Item Difficulty

- Clear that some items easy and some hard
- Assuming all same leads to bad marginal fit
- Hard to estimate even with 10 annotators/item
  - Posterior intervals too wide

Modeling Item Difficulty

- Logistic Item-Response models with shape used in social sciences (e.g. education and voting)
- Use logistic scale (maps $(-\infty, \infty)$ to $[0, 1]$)
- α_j: annotator j's bias (ideally 0)
- δ_j: annotator j's discriminativeness (ideally ∞)
- β_i: item i's "location" plus "difficulty"
- $x_i \sim \text{logit}^{-1}(\delta_j(\alpha_i - \beta_j))$

Modeling Item Difficulty (Cont.)

- Place normal (or other) priors on coefficients,
  e.g. $\beta_i \sim \textsf{Norm}(0, \sigma^2), \quad \sigma^2 \sim \textsf{Unif}(0, 100)$
- Priors may be estimated as before; leads to pooling of item difficulties.
- Need more than 5-10 coders/item for tight posterior on difficulties
- Model has better χ^2 fits, but many more params
- Harder to estimate computationally in BUGS
- Full details and code in paper

Extensions

Extending Coding Types

- Multinomial responses (Dirichlet-multinomial)
- Ordinal responses (ordinal logistic model)
- Scalar responses (continuos responses)

Active Learning

- Choose most useful items to code next
- Typically balancing two criteria
  - high uncertainty
  - high typicality (how to measure?)
- Can get away with fewer coders/item
- May introduce sampling bias
- Compare supervision for high certainty items
  - High precision (for most customers)
  - High recall (defense analysts and biologists)

Code-a-Little, Learn-a-Little

- Semi-automated coding
- System suggests labels
- Coders correct labels
- Much faster coding
- But may introduce bias

- Hugely helpful in practice

Probabilistic Training and Testing

- Use probabilistic item posteriors for training
- Use probabilistic item posteriors for testing
- Directly with most probabilistic models (e.g. logistic regression, multinomial)
- Or, train/test with posterior samples
- Penalizes overconfidence of estimators (in log loss)
- Demonstrated theoretical effectiveness (Smyth et al.)
- Need to test in practice

Semi-Supervised Models

- Easy to add in supervised cases with Bayesian models
  - Gibbs sampling skips sampling for supervised cases
- May go half way by mixing in “gold standard” annotators
  - Fixed high, but non-100% accuracies
  - Stronger high accuracy prior

Multimodal (Mixture) Priors

- Model Mechanical Turk as mixture of spammers and hammers
- This is what the Mechanical Turk data suggests
- May also model covariance of sensitivity/specificity

Annotator and Item Random Effects

- May add random effects for annotators
  - amount of annotator training
  - number of items annotated
  - annotator native language
  - annotator field of expertise
- Also for Items
  - difficulty (already discussed)
  - type of item being annotated
  - frequency of item in a large corpus

Jointly Estimate Model and Annotations

- Can train a model with inferred (probabilistic) gold standard
- Can use trained model like another annotator
- Raykar, Vikas C., Shipeng Yu, Linda H. Zhao, Anna Jerebko, Charles Florin, Gerardo Hermosillo Valadez, Luca Bogoni, and Linda Moy. 2009. Supervised Learning from Multiple Experts: Whom to trust when everyone lies a bit. In *ICML*.

Bayesian κ Estimates

- Calculate expected κ for two annotators
- Calculate expected κ for two new annotators from pool
- Calcluate confidence/posterior uncertainty of κ
  - Could estimate confidence intervals for κ w/o model

The End

- References
  - `http://lingpipe-blog.com/`
- Contact
  - `carp@alias-i.com`
- R/BUGS (Anon) Subversion Repository

  ```
  svn co https://aliasi.devguard.com/svn/sandbox/hierAnno
  ```