Probabilistic n-Choose-k Models for Classification and Ranking

Kevin Swersky[1], Daniel Tarlow[1], Ryan P. Adams[2], Richard S. Zemel[1], and Brendan J. Frey[1]

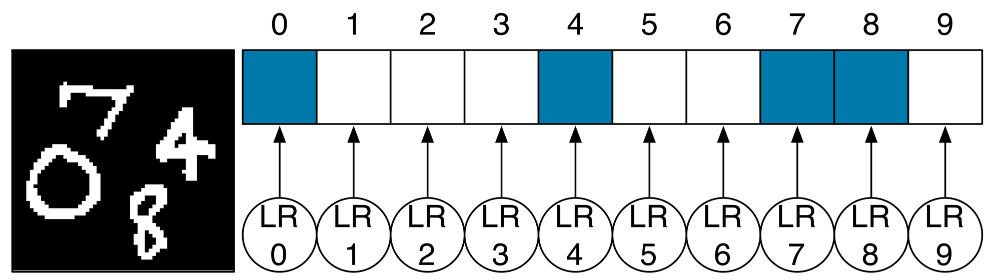
1. University of Toronto 2. Harvard University

Problem and Motivation

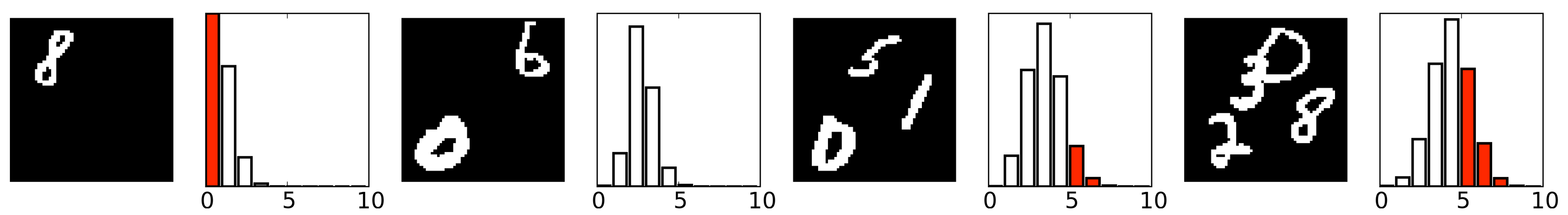
Multi-label classification: predict multiple outputs, e.g., identify multiple objects in an image.

Often desirable to model count structure. E.g., in order to identify *which* objects there are, it is helpful to know *how many* there are.

- Simple idea: multiple independent logistic regression (LR) classifiers.

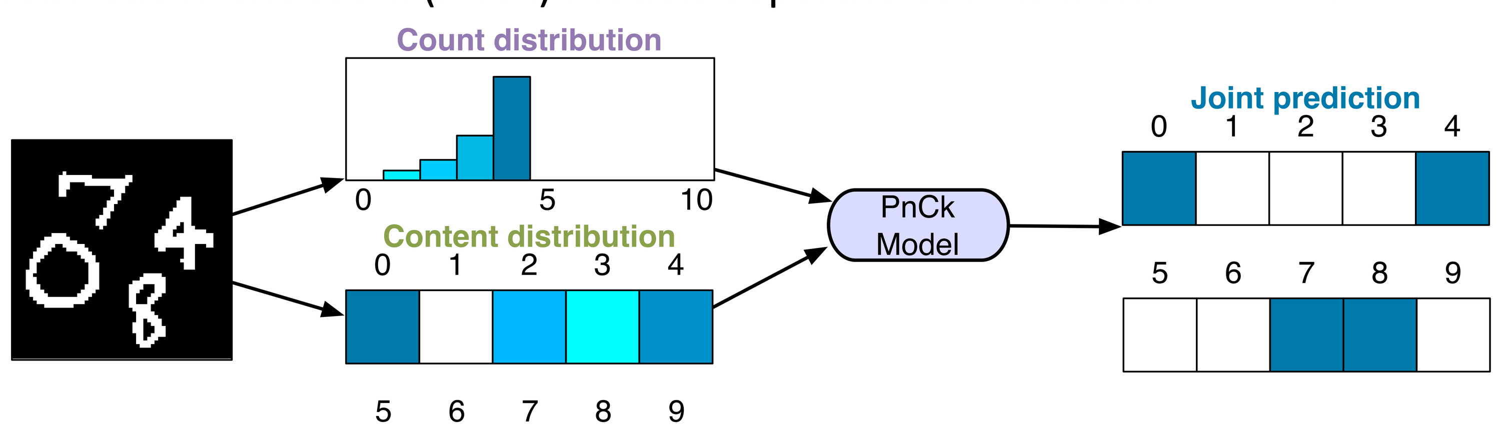


- Problem: LR uses the same parameters to both *identify* and *count* objects.
- Even though there are always 1 to 4 objects in each image, Logistic regression may predict 0 objects, 5 objects, 6 objects, etc., this limits its modeling ability.



Examples from the Embedded MNIST dataset and distribution over counts learned by LR.

- Probabilistic n-Choose-k (PnCk) models separate counts from content.



- The count distribution can be input-dependent, or given as a prior.

The Binary n-Choose-k Model (BnCk)

Setting: learn to predict multiple binary outputs.

Features	Parameters	Model Inputs	Outputs
$\boldsymbol{x}$	$\mathbf{W}$	$\boldsymbol{\theta} = \mathbf{W}\boldsymbol{x}$	$\boldsymbol{y} \in \{1, 2, \ldots, R\}^D$

- Define a subset of the variables by $c \subseteq \{1, \ldots, D\}$ and complement $\bar{c} = \{1, \ldots, D\} \setminus c$.

- Draw k from a prior distribution $p(k)$ over counts k.
- Draw k variables to take on label 1, where the probability of choosing subset c is given by

$$p(\boldsymbol{y}_c = \mathbf{1}, \boldsymbol{y}_{\bar{c}} = \mathbf{0} \mid k) = \begin{cases} \frac{\exp\{\sum_{d \in c} \theta_d\}}{Z_k(\boldsymbol{\theta})} & \text{if } |c| = k \\ 0 & \text{otherwise} \end{cases}$$

Connection to Logistic Regression

- Multiple output logistic regression can be viewed as a binary n-choose-k model.
- Let $Z_k(\boldsymbol{\theta}) = \sum_{\boldsymbol{y}, |\boldsymbol{y}|=k} \exp(\sum_d \theta_d y_d)$, $Z(\boldsymbol{\theta}) = \sum_k Z_k(\boldsymbol{\theta})$, and assume $\sum_d y_d = k$,

$$p(\boldsymbol{y}, k; \boldsymbol{\theta}) = p(k; \boldsymbol{\theta}) p(\boldsymbol{y} \mid k; \boldsymbol{\theta}) = \frac{Z_k(\boldsymbol{\theta})}{Z(\boldsymbol{\theta})} \frac{\exp\{\sum_{d \in c} \theta_d\}}{Z_k(\boldsymbol{\theta})} = \prod_d \frac{\exp\{\theta_d y_d\}}{1 + \exp\{\theta_d\}}$$

- Logistic regression *implicitly* models counts using the "prior" $p(k; \boldsymbol{\theta}) = \frac{Z_k(\boldsymbol{\theta})}{Z(\boldsymbol{\theta})}$
  - Induces independence between output variables.
- The prior distribution is called a Poisson-Binomial distribution (Chen et al., 1994).
  - Distribution over the number of successes in independent Bernoulli trials with *different* probabilities.

The Ordinal n-Choose-k Model (OnCk)

Setting: given a set of items and associated relevance scores, learn to rank the items.

- Given initial set of unlabeled variables $\boldsymbol{y}^u = \boldsymbol{y}$, let k_r be the number of variables with label r, and $\boldsymbol{k} = (k_1, \ldots, k_R)$, such that $\sum_r k_r = D$.

- Sample relevance score counts $k_R, \ldots, k_1$ jointly from $p(\boldsymbol{k})$.
- Repeat for $r = R$ to 1:
  - Choose a subset c_r of k_r unlabeled variables from $\boldsymbol{y}^u$ and assign them relevance label r. Choose subsets with probability:

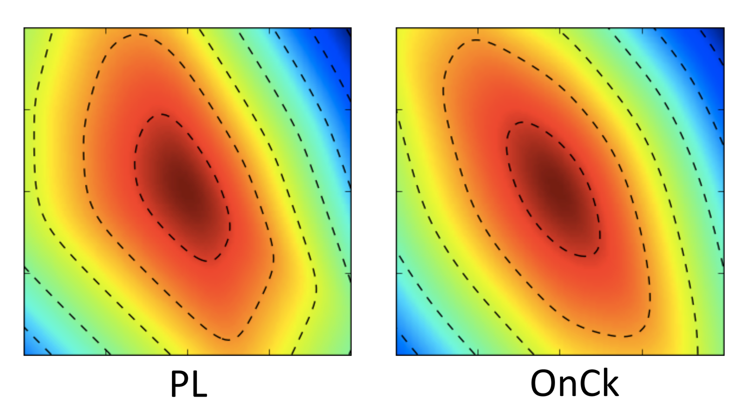
$$p(\boldsymbol{y}^u_{c_r} = \mathbf{1}, \boldsymbol{y}^u_{\bar{c}_r} = \mathbf{0} \mid k_r) = \begin{cases} \frac{\exp\{\sum_{d \in c_r} \theta_d\}}{Z^u_k(\boldsymbol{\theta})} & \text{if } |c_r| = k_r \\ 0 & \text{otherwise} \end{cases},$$

  where $Z^u_k(\boldsymbol{\theta})$ is a sum over all subsets of size k_r from $\boldsymbol{y}^u$.
  - Remove $\boldsymbol{y}^u_{c_r}$ from $\boldsymbol{y}^u$.

- Training objective is *convex*, only tuning parameter is L2 penalty strength.
- At test-time, the quality of a ranking is evaluated using a *gain function*, e.g., NDCG, Precision@K. Finding the optimal ranking under OnCk is trivial for these measures.

Theorem 1. *Under an ordinal n-choose-k model, the optimal decision theoretic predictions for monotonic ranking gains, such as NDCG and Precision@K, are made by sorting the* θ *scores.*

- OnCk can be viewed as a generalization of the Plackett-Luce (PL) distribution.
- With PL, $R = D$; we draw *one* item at a time. With OnCk we draw *groups* of items.
- Both distributions yield empirically similar objective functions.
- OnCk training is *efficient* and *exact*.

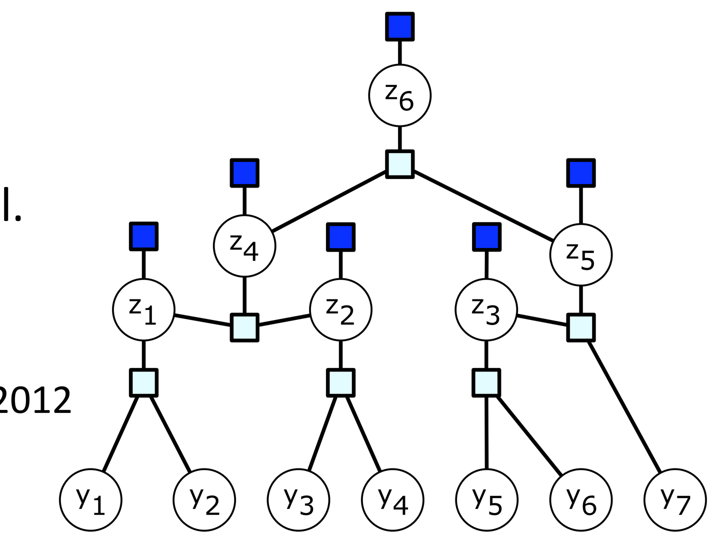


Objective comparison for a synthetic dataset with 2D inputs.

Efficient Likelihood Computation

- The PnCk count-conditional likelihood involves summing over all subsets of of size k.
- Can be viewed as a Markov Random Field with unary potentials and a global cardinality potential.
- Can efficiently compute the count conditional likelihood in $O(D \log^2 D)$.

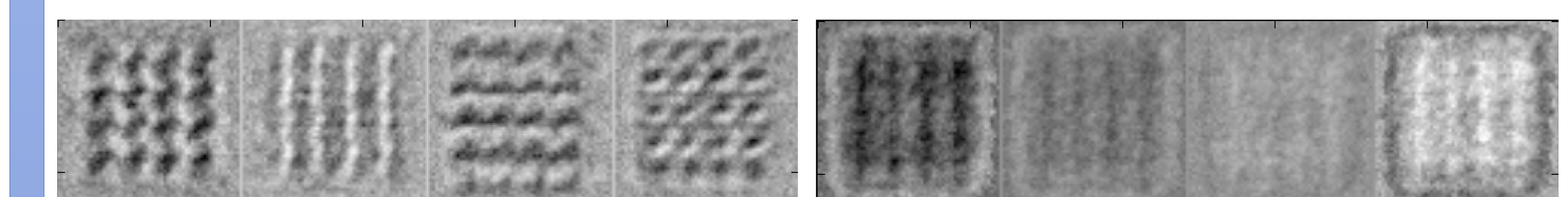
Belfore, 1995, Tarlow et al., 2012



Experiments

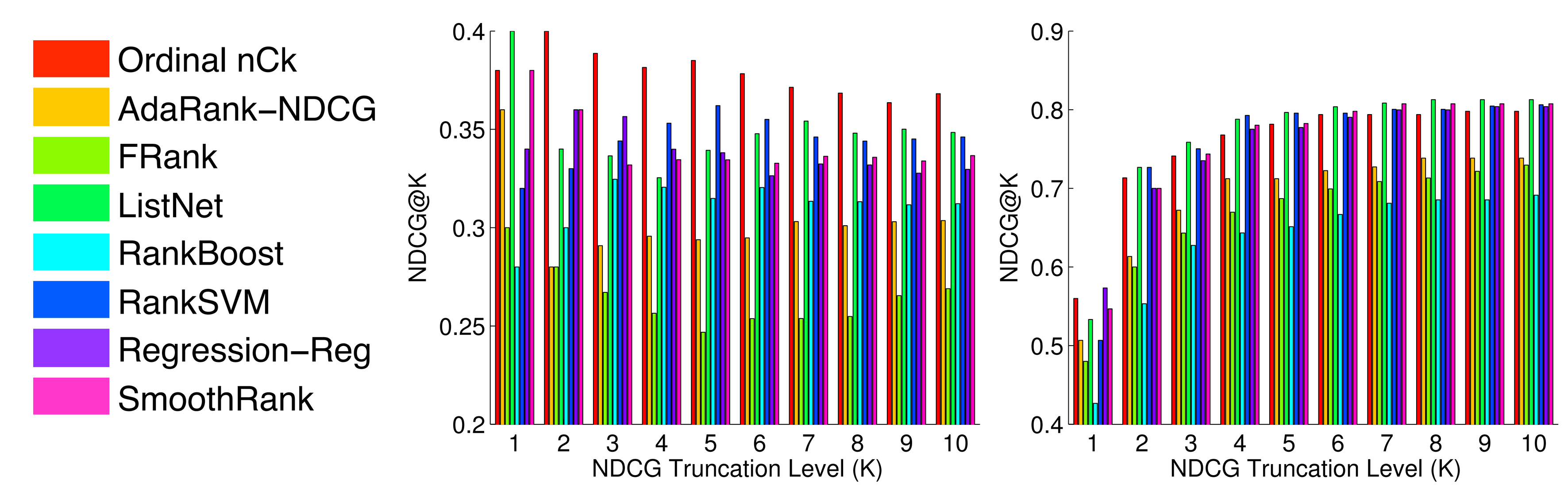
Modeling the number of objects in an image.

- We train a BnCk model with an input-dependent count prior $P(k|x)$.
- Separates *which* digits appear in an image from *how many*.
- LR test set log-likelihood: -2.84. BnCk test set log-likelihood: -1.95.



Ranking with weak labels on the LETOR 3.0 datasets.

- Each input is a query with multiple documents and associated relevance scores.
- Output is a ranking of the documents within each query.



Top-k Classification

- The inputs are an image and a single ground-truth label, the outputs are the top k predictions of the model.
- We train to maximize the expected accuracy under a top-k evaluation criterion.
- Overfitting can be an issue, but training and testing with top-k is promising.

Evaluation Criterion

Training objective	Strong L2 penalty: Top 1 / Top 3 / Top 5	Weak L2 penalty: Top 1 / Top 3 / Top 5
LR	0.606 / 0.785 / 0.812	0.545 / 0.716 / 0.766
Top 1	0.621 / 0.796 / 0.831	0.574 / 0.755 / 0.804
Top 3	0.614 / 0.792 / 0.834	0.558 / 0.771 / 0.813
Top 5	0.602 / 0.787 / 0.834	0.523 / 0.767 / 0.823

Top 1 is equivalent to softmax regression

Training Accuracy on Caltech 101 Silhouettes.