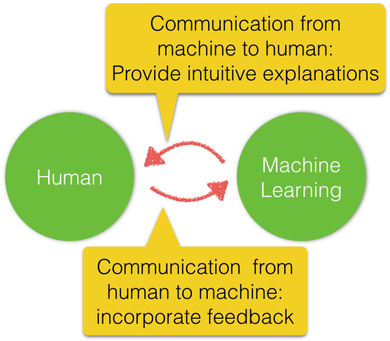
Bayesian Case Model (BCM): Generative Approach for Case-based Reasoning and Prototype Classification

Been Kim, Cynthia Rudin and Julie Shah
{beenkim, rudin, julie_a_shah}@csail.mit.edu

Massachusetts Institute of Technology

Motivation

Develop interactive machine learning framework to bring together data, machine learning algorithms and human experts' knowledge.



First step: communication from machine to human

Why Case-based Reasoning?

1. Case-based Reasoning (CBR) is intuitive.

Example of CBR (simplified)
a. My child is sick. He is having fever, nausea and coughs.
b. Find the illness that has the most similar symptoms ➔Flu
c. Apply the solution for flu: rest, ibuprofen, liquid.

2. Exemplar-based reasoning is the way humans think – machine can better support peoples' decision-making by representing data in the same way.

Human's tactical decision is based on exemplar-based reasoning.[1]
Skilled fire fighters use recognition-primed decision making — a situation is matched to typical cases. [2,3]


Limitations of CBR

- Does not leverage data
- Requires previous solutions
- Does not scale to complex problems

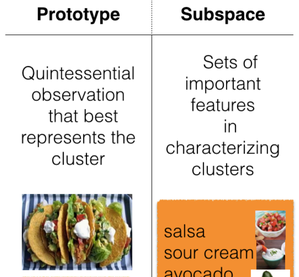
[1] M.S. Cohen, J.T. Freeman, and S. Wolf. Metarecognition in time-stressed decision making: Recognizing, critiquing, and correcting. Human Factors, 1996.
[2] A. Newell and H.A. Simon. Human problem solving. Prentice-Hall Englewood Cliffs, 1972
[3] G.A. Klein. Do decision biases explain too much. HFES, 1989.

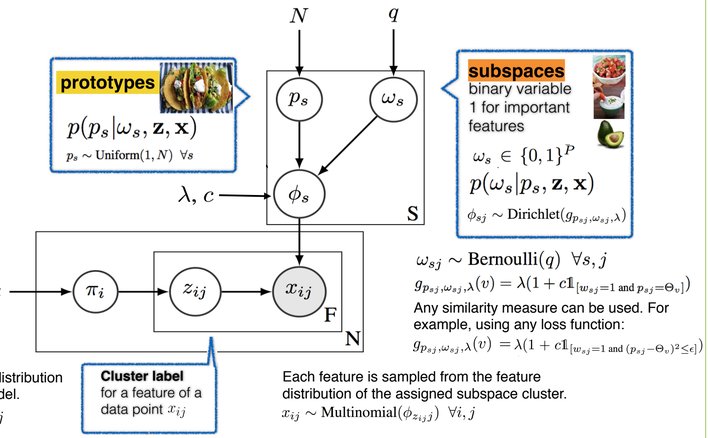
Bayesian Case Model (BCM)

Leverage the power of examples (prototypes) and hot features (subspaces) to explain machine learning results

Bayesian generative models + Case-based reasoning = Bayesian Case Model

Prototype	Subspace
Quintessential observation that best represents the cluster	Sets of important features in characterizing clusters





We model the underlying distribution using a vanilla mixture model.
$z_{ij} \sim \text{Multinomial}(\pi_i)\ \forall i, j$
$\pi_i \sim \text{Dirichlet}(\alpha)\ \forall i$

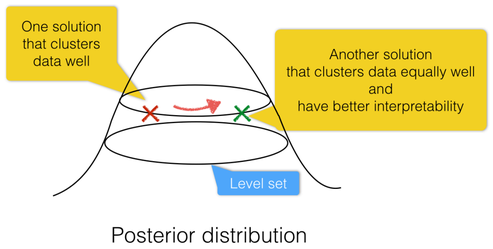
Cluster label for a feature of a data point x_{ij}

$\omega_{sj} \sim \text{Bernoulli}(q)\ \forall s, j$
$g_{p_{sj},\omega_{sj},\lambda}(v) = \lambda(1 + c\mathbb{1}_{[w_{sj}=1 \text{ and } p_{sj}=\Theta_v]})$
Any similarity measure can be used. For example, using any loss function:
$g_{p_{sj},\omega_{sj},\lambda}(v) = \lambda(1 + c\mathbb{1}_{[w_{sj}=1 \text{ and } (p_{sj}-\Theta_v)^2 \le \epsilon]})$

Each feature is sampled from the feature distribution of the assigned subspace cluster.
$x_{ij} \sim \text{Multinomial}(\phi_{z_{ij}j})\ \forall i, j$

Improving interpretability without sacrificing performance

Joint inference on prototypes, subspaces and cluster labels is the key.



Posterior distribution

Collapsed Gibbs sampling for inference

Integrating out and for efficient inference

$$p(z_{ij} = s | z_{i\neg j}, \mathbf{x}, p, \omega, \alpha, \lambda) \propto \frac{\alpha/S + n_{(s,i,\neg j,\cdot)}}{\alpha + n} \times \frac{g(p_{sj}, \omega_{sj}, \lambda) + n_{(s,\cdot,j,x_{ij})}}{\sum_s g(p_{sj}, \omega_{sj}, \lambda) + n_{(s,\cdot,j,\cdot)}}$$

where $n_{(s,i,j,v)} = \mathbb{1}(z_{ij} = s, x_{ij} = v)$

$$p(\omega_{sj} = b | q, p_{sj}, \lambda, \phi, \mathbf{x}, \mathbf{z}, \alpha) \propto \begin{cases} q \times \frac{\mathbf{B}(g(p_{sj},1,\lambda) + n_{(s,\cdot,j,\cdot)})}{\mathbf{B}(g(p_{sj},1,\lambda))} & \text{b = 1} \\ 1-q \times \frac{\mathbf{B}(g(p_{sj},0,\lambda) + n_{(s,\cdot,j,\cdot)})}{\mathbf{B}(g(p_{sj},0,\lambda))} & \text{b = 0} \end{cases}$$

where $\mathbf{B}$ is the Beta function

An example of BCM

prototypes
subspaces

Raw data

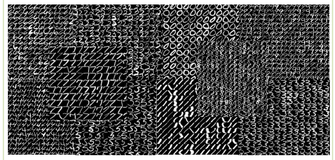
Results

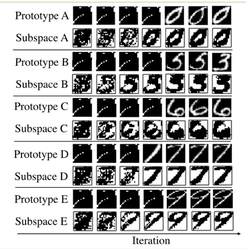
1. Learned prototypes and subspaces

Dataset	Prototypes and Subspaces from BCM
a. Recipes Data from computer cooking contest: liris/cnrs.fr/ccc/ccc2014	

Prototypes from BCM (Recipe names)	Subspaces from BCM and Ingredients (subspaces)
Herbs and tomato in pasta	basil, garlic, Italian seasoning, oil, pasta, pepper, salt, tomato
Generic chili recipe	beer, chili powder, cumin, garlic, meat, oil, onion, pepper, salt, tomato
Microwave brownies	baking powder, sugar, butter, chocolate, chopped pecans, eggs, flour, salt, vanilla
Spiced-punch	cinnamon stick, lemon juice, orange juice, pineapple juice, sugar, water, whole cloves

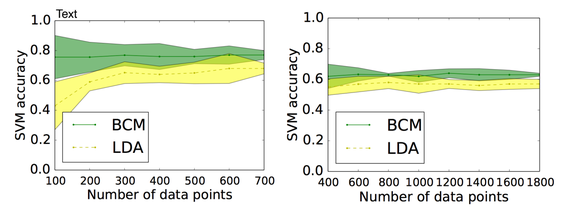
b. USPS handwritten digits
http://www.cs.nyu.edu/~roweis/data.html





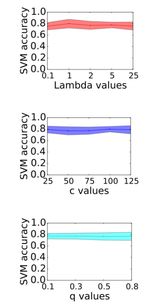
2. BCM maintains accuracy while achieving interpretability

Evaluated using features learned by BCM or LDA + SVM



Handwritten digit dataset

20Newsgroups dataset



Sensitivity analysis

3. BCM improves humans understanding

- Objective measure of human understanding
- Participant's task is to assign the ingredients of a specific dish (a new data point) to a cluster.
- Statistically significantly better performance with BCM **85.9% v.s. 71.3%**

$$p \ll 0.001$$