T-122.102 Yhteisesiintymädatan analysointi [T-122.102] (1)

Kai Puolamäki

20 January 2004
(corrected version)

t122102@james.hut.fi

http://www.cis.hut.fi/Opinnot/T-122.102/

References (PDFs can be found by using Google) [T-122.102] (2)

N. Tisby, F. C. Pereira, W. Bialek, *The Information Bottleneck Method*, 1999.

R. Bekkerman, R. El-Yaniv, N. Tisby, Y. Winter, *Distributional Word Clusters vs. Words for Text Categorization*, 2003.

J. Sinkkonen, S. Kaski, *Clustering based on conditional distributions in an auxiliary space*, 2002.

J. K. Pritchard, M. Stephens, P. Donnelly, *Inference of Population Structure Using Multilocus Genotype Data*, 2000.

D. M. Blei, A. Y. Ng, M. I. Jordan, *Latent Dirichlet Allocation*, 2002.

W. Buntine, *Variational Extensions to EM and Multinomial PCA*, 2002.

D. M. Blei, M. I. Jordan, *Modeling Annotated Data*, 2003.

T. S. Jaakkola, D. Haussler, *Exploiting generative models in discriminative classifiers*, 1998.

J. Lafferty, G. Lebanon, *Information diffusion kernels*, 2002.

References (continued) [T-122.102] (3)

M. Seeger, *Covariance Kernels from Bayesian Generative Models*, 2001.

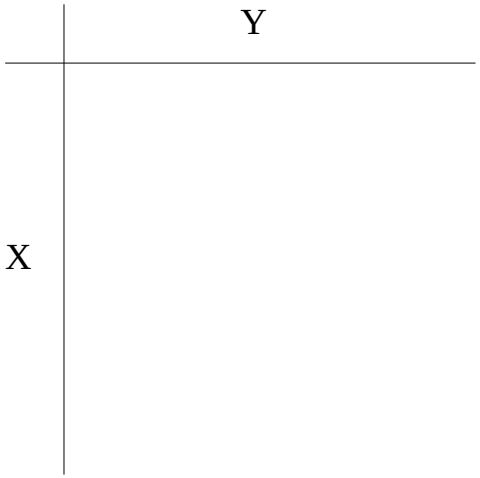
H. Steck, T. S. Jaakkola, *(Semi-)Predictive Discretization During Model Selection*, 2003.

About the problem [T-122.102] (4)

- Themes:
  - Discrete data
  - Joint distributions
  - Principled approaches (e.g. mutual information, generative models)

Joint distributions [T-122.102] (5)

- Variables: X (e.g. documents) and Y (e.g. word counts)
- Clustering: $X \to \tilde{X}$ (e.g. document clusters) and/or $Y \to \tilde{Y}$ (e.g. topics)



Various solutions [T-122.102] (6)

- Mutual information -motivated methods (e.g. IB, DC)
- Generative models (e.g. LDA, mPCA)
- Kernel methods (e.g. Fischer kernels)
- ...
- Discretizing data

Mutual information [Mutual information] (7)

- Mutual information:

$$I(X,Y) = \sum_{x \in X, y \in Y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)}$$

- $I(X,Y)$ is the Kullback Leibler divergence, $KL(P,Q)$, between the joint distribution $P(X,Y)$ and the distribution that assumes that X and Y are statistically independent, $Q(x,y) = P(x)P(y)$. If we are picking samples from distribution P then the mutual information (or KL divergence) measures the average amount of information the samples give for deciding that the samples are not from the distribution Q.
- General idea: find an intermediate representation that maximizes $I(X,Y)$.

Information bottleneck (IB) [Mutual information] (8)

- Shannon's theory: X is transmitted using coding defined by $\tilde{X}$ and then decoded to Y,b

$$X \to \tilde{X} \to Y \quad .$$

  What is the optimal coding (maximizes $I(X,Y)$ for a given channel capacity)?
- Equivalent(?) problem: find optimal coding/clusterng X that transmits optimal value of information for a fixed value of $I(\tilde{X}|X)$. I.e., maximize

$$I(\tilde{X},Y) - \beta I(\tilde{X},X) \quad ,$$

  where β is a Lagrange multiplier.
- Works only for discrete data ($I(\tilde{X},X)$?)
- No need for generative model

Clustering text [Mutual information] (9)

- Input:
  - X (documents), Y (word counts in documents)
  - Empirical joint distribution, $P(X,Y)$
  - Number of clusters, k
  - Parameters for the optimization algorithm
- Output:
  - Cluster assignement probabilities for the documents ("coding"), $P(\tilde{X}|X)$
  - Cluster centroids ("decoding"), $P(Y|\tilde{x}_i)$, $i = 1,\ldots,k$.

Clustering text (continued) [Mutual information] (10)

[Inserts figure 2 and 4 from Tishby et. al 2003 here]

[PS. Our scanner is currently unavailable...]

Discriminative Clustering (DC): a very cursory introduction [Mutual information] (11)

- See e.g. Sinkkonen, Kaski, 2002.
- Clustering algorithm based on KL divergence. Cost function:

$$\sum_j \int dx y_j(x) KL(p(c|x), \Psi_j) p(x) \quad ,$$

  where Ψ_j is a cluster prototybe and $y_j(x)$, $\sum_j y_j(x) = 1$, is a cluster membership function.
- Inputs are pairs of (x_k, y_k) data, where $x_k \in R^n$ (data, $\sim X$) and $c_k \in \{1,...,K\}$ (class, $\sim Y$).
- Cost function is essentially equivalent to $I(\tilde{X}, Y)$
- $I(\tilde{X}, X)$ is fixed by the model complexity and the number of clusters

Generative models [Generative models] (12)

- Assume that data has been generated by some model having parameters θ
- Idea: find the parameters θ by optimizing $\max_\theta P(X|\theta)$, $\max_\theta P(\theta|X)$ or $P(\theta|X)$.
- Latent Dirichlet Allocation (LDA) and Multinomial PCA (mPCA) are essentially the same thing (some differences in priors and optimization algorithms)

The basic model [Generative models] (13)

- Each individual/document, indexed by $d = 1, \ldots, D$, is composed of N_d loci/words (L different alleles/words). Each allele/word originates from one of k populations/topics, denoted by $z_n = 1, \ldots, k$, $n = 1, \ldots, N_d$:

$$\begin{aligned} P(\theta|\alpha) &\sim Dirichlet(\alpha; k) \quad , \\ P(z_n|\theta) = \theta_{z_n} &\sim Multinomial(\theta; k) \quad , \\ P(x_n|z_n, \beta) = \beta_{x_n, z_n} &\sim Multinomial(\beta\theta; L) \quad . \end{aligned}$$

- Generative model (comp. graphical model):

$$P(X|\alpha) = \int d\theta P(\theta|\alpha) \prod_{n=1}^{N_d} \sum_{z_n=1}^{k} p(x_n|z_n, \beta) p(z_n|\theta)$$

- Non-negative generalization of PCA
  ($Gaussian \leftrightarrow Dirichlet$, $Multinomial$)

MCMC approach to LDA/mPCA [Generative models] (14)

- Pritchard et al. 2000, unreferenced by the (first) LDA/mPCA papers
- The posterior probability distribution $P(\theta, \beta|X)$ is sampled using a Markov Chain Monte Carlo (MCMC) method
- Input:
  - The genomes of individuals
  - Number of clusters, k
- Output:
  - The population decompositions for each individual, $\{\theta_i\}_{i=1,\ldots,k}$
  - The mixing matrix $P(x_n|z_n, \beta) = \beta_{x_n, z_n} \in R^{L \times k}$,

MCMC approach to LDA/mPCA [Generative models] (15)

[Insert figure 4 from Pritchard et al. 2000 here]

LDA/mPCA [Generative models] (16)

- $\max_\lambda P(X|\lambda)$, where λ denote the parameters, is found using variational extension to EM
- Faster than MCMC but may also be more unstable
- Variational means that the joint distribution of "hidden variables" $h \sim \theta, z$ is approximated by a product distribution $q(\theta, z|\lambda') = q(\theta) \prod_n q(z_n)$, having some parameters λ'.
- Cost function:

$$\begin{aligned} \mathcal{L}(\lambda, \lambda') &= \log P(X|\lambda) - KL(q(h|\lambda'), p(h|X, \lambda) \\ &= H(q(h|\lambda')) + E_{q(h|\lambda')}\{\log P(X, h|\lambda)\} \quad . \end{aligned} \tag{1}$$

  The cost function is maximized by iteratively minimizing the KL divergence with respect to λ' and maximizing the expectation with respect to λ, resulting to a lower bound to $\log P(X|\lambda)$.

Extension: Annotations [Generative models] (17)

- Each “document” consists of a image data (feature vectors) and captions
- Generative model: [Insert figure 2 from Blei, Jordan, 2003 here]
- Variational approximation
- Automatic annotations: [Insert figure from 5 Blei, Jordan, 2003 here]

Mutual information kernels [Kernel methods] (18)

- See e.g. Seeger, 2001
- Assume the data set X is generated by some model, having parameters θ
- Then the joint distribution of two variables $x_i \in X$ can be written as

$$P(x_1,x_2) = \int d\theta P(\theta|X)P(x_1|\theta)P(x_2|\theta) \quad ,$$

and the marginal distribution as

$$P(x) = \int d\theta P(\theta|X)P(x|\theta) \quad .$$

- Define sample mutual information by

$$I(x_1,x_2) = \log \frac{P(x_1,x_2)}{P(x_1)P(x_2)} \quad . \qquad (2)$$

Mutual information kernels (continued) [Kernel methods] (19)

- Kernel should at least satisfy $K(x_1,x_2) = K(x_2,x_1) \geq 0$.
- Intuitive idea: if $K(x_1,x_2)$ is large then x_1 and x_2 are similar and if $K(x_1,x_2)$ is small then x_1 and x_2 are dissimilar. Kernel thus defines a similarity measure.
- $I(x_1,x_2)$ defines some kind of a similarity measure. However, it is not always positive and it is thus not a good kernel.

Mutual information kernels (continued) [Kernel methods] (20)

- Define a positive definite kernel by exponential embedding,

$$K(x_1,x_2) = \exp\left(-D^2(x_1,x_2)\right) \quad ,$$

where

$$D^2(x_1,x_2) = I(x_1,x_1) + I(x_2,x_2) - 2I(x_1,x_2) \quad .$$

- If the kernel can be represented as an inner product in Euclidean space, $I(x_1,x_2) = \phi_{x_1}^T \phi_{x_2}$, then the distance $D^2(x_1,x_2)$ corresponds to the squared Euclidean distance $|\phi_{x_1} - \phi_{x_2}|^2$.
- Classification task (two classes, $S_i = \pm 1$). Optimize e.g. discriminant:

$$\mathcal{L}(x) = \sum_i S_i \lambda_i K(x,x_i) \quad .$$

Fischer kernels [Kernel methods] (21)

- Approximate $P(\theta|X)$ with a Gaussian around MAP parameters, $\hat{\theta}$,

$$P(\theta|X) \propto \exp\left(-\frac{1}{2}(\theta-\hat{\theta})^T H^{-1}(\theta-\hat{\theta})\right) \quad ,$$

  and make linear approximation of $P(x_i|\theta)$,

$$\log P(x_i|\theta) \approx \log P(x_i|\hat{\theta}) + U_{x_i}(\theta-\hat{\theta}) \quad ,$$

  where $U_{x_i} = \nabla_{\hat{\theta}} \log P(x_i|\hat{\theta})$ (Fischer score).

- This results to the original Fischer kernel (Jaakkola, Haussler, 1998),

$$I(x_1,x_2) = U_{x_1}^T H^{-1} U_{x_2}$$

  and

$$K(x_1,x_2) = \exp\left(-(U_{x_1}-U_{x_2})^T H^{-1}(U_{x_1}-U_{x_2})\right) \quad .$$

Other approaches [Kernel methods] (22)

- Exact kernels, without Gaussian approximation

[Insert figure 1 from Lafferty et al. here]

Discretization [Others] (23)

- How to discretize continuous data optimally?
- The discretization may affect the results significantly
- E.g. Steck, Jaakkola, 2003 (find discretization that maximizes the likelihood of the data)

Summarizing [T-122.102] (24)

- Themes:
  - Discrete data
  - Joint distributions
  - Principled approaches (e.g. mutual information, generative models)
- Recent advances:
  - Mutual information based methods (e.g. IB)
  - Generative model based methods (e.g. LDA/mPCA)
  - Information theoretical kernels (concepts from mutual information and generative models)
  - Discretizing data
  - Other stuff?