Exploiting generative models in discriminative classifiers

By
Jaakkola
Haussler

Outline of paper

- Generative probability models deal with missing information and variable length sequences.
- Discriminative methods perform superior to probability models.
- The author tries to develop an ideal classifier which combines both the approaches by deriving kernels function

Introduction

- Speech, vision, text etc. are difficult to deal in statistical classification problem
- Problem is no systematic way to get relationship between examples
- We propose a general method for extracting discriminatory features and these features are more suited to kernel methods

Kernel methods

- With a training set examples X_i and corresponding labels S_i
- In kernel methods, the label for a new example is determined by the weighted sum of the training labels
- Weighting consists of
  1.overall importance of the example X_i represented by λ_i
  2.measure of pairwise "similarity" between the X_i and X expressed in terms of K (X_i , X)

Kernel methods cont..

- The predicted label $\hat{S}$ for the new example is derived from
  $\hat{S} = \text{sign} (\Sigma_i S_i \lambda_i K (X_i , X))$
- To say it a kernel method, two things are to be clarified
  1.classification loss
  2.the choice of kernel function

Generalized linear models

- Probability of the label S is
  $P(S|X, \theta) = \sigma (S\, \theta^T X)$
  where $\sigma(Z) = (1+e^{-z})^{-1}$
- The maximum a posteriori estimate for the parameters θ given a training set of examples is found by maximizing the following penalized log-likelihood

$\Sigma_i \log P(S_i|X_i, \theta) + \log P(\theta) = \Sigma_i \log \sigma(S_i \theta^T X_i) - 1/2\theta^T \Sigma^{-1} \theta + c$
here the constant c doesn't depend on θ

Generalized linear models cont..

- The solution to this problem can be
  $\theta = \Sigma_i S_i \lambda_i \Sigma X_i$ where
  $\partial/\partial Z \log \sigma(Z)|_{z=s_i \theta^T X_i} = \sigma(-S_i \theta^T X_i)$
- This can be put back into conditional probability model gives
  $P(S|X, \theta) = \sigma(S \Sigma_i S_i \lambda_i (X_i^T \Sigma X))$
- Here we can identify $K(X_i, X) = X_i^T \Sigma X$

Kernel function

- For a kernel function to be valid it should positive semi-definite
- According to Mercers theorem,
  $K(X_i, X) = Ø^T_{Xi} Ø_{Xj}$
- Specifying a simple inner product in the feature space defines a Euclidean metric space

Kernel function cont..

- Euclidean distances between feature vectors is calculated as
  $|| Ø_{Xi} - Ø_{Xj} ||^2 = K(X_i, X_i) - 2K(X_i, X_j) + K(X_j, X_j)$
- It also defines a pseudo metric in the original example space
- Thus the kernel embodies prior assumptions about the metric relations between the original examples

The fisher kernel

- Attempt to find natural comparison between examples induced by the generative model
- Use gradient space to capture the generative process
- Gradient of likelihood describes the process of generating particular eaxmple

Fisher kernel cont..

- Consider a parametric class of $P(X|\theta)$, where $\theta \in \Theta$.
- Defines a Riemannian manifold M_Θ with a local metric given by the Fisher information matrix I where
  $I = Ex\{UxUx^T\}$, $Ux = {}_\theta \log P(X|\theta)$.
- Ux is called the fisher score
- The local metric on M_Θ defines a distance between the current model $P(X|\theta)$ and a nearby model $P(X|\theta+\delta)$

Fisher kernel cont..

- The distance is given by
  $D(\theta, \theta+\delta) = ½ \delta^T I \delta$
- The fisher score mapping
- Gradient Ux used to define the direction of steepest ascent along the manifold

Fisher kernel cont..

- From metric point of view a scaled/translated kernel $K(X_i X_j) = cK(X_i X_j) + c_0$ where $c_0>0$
- here c relates to the overall priori variance of remaining parameters
- The fisher kernel provides only the basic comparison between the examples defining what is meant by an "inner product"

Fisher kernel cont..

- Using fisher kernel, a linearly separable hyper plane in the feature space
- Examples may not linearly separable
- Transforming the fisher kernel according to $K^{\sim}(X_i X_j)=1+(K(X_i X_j))^m$ and using the resulting as a classifier

Properties of kernel function

- For any probability model $P(X|\theta)$ with parameters θ, the fisher kernel $K(X_i X_j) = U^T_{xi} I^{-1} U_{xj}$ where U_{xi} has the following properties
  1. it is a valid kernel function
  2. it is invariant to any invertible transformation of parameters
  3. a kernel classifier employing the fisher kernel derived from a model that contains a label as a latent variable

Properties cont..

- The first property is positive definite
- Kernel was defined with reference only to the manifold M_Θ
- Third property can be established based on basis of discriminative derivation of this kernel

Experiments

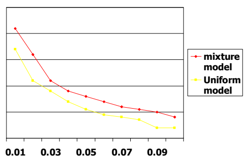
- Consider two examples DNA and recognition of remote homologies between protein sequences
- Consider 9350 DNA fragments
- 2029 are true examples in a sequence X over the DNA alphabet {A,G,T,C} of length 25 centered on the consensus 'GT' at the 5' splice boundary

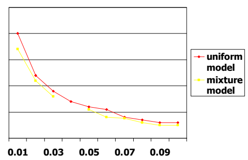
Experiment1

- The rest are false examples similar sequences centered at 'GT' but not near 5'splice sites
- We test performance of the combined classifier on the quality of underlying model
- The model chosen here is

  $P(X|\theta)=\prod^{25}_{i=1} p(X_i|\theta_i)$

Result 1

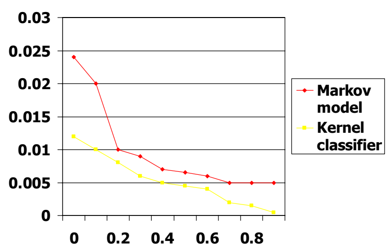




Experiment 2

- Problem of recognizing remote homologies between protein sequences that have low residue identity
- A lot of recent has been done in refining hidden Markov models
- Picked a particular super family and left out one of 4 major families in this super family.
- This gave us the scheme for 4-fold cross validation

Result 2



Conclusion

- Kernel functions derived provides a mechanism for incorporating generative models into discriminative classifiers.
- The power of new classifier is use of fisher scores as features in place of original examples.

Exploiting Generative Models in Discriminative Classifiers

Overview from Kai Xu

OVERVIEW

- What this paper deals with
  - To predict the label for a new example, it's very important to have a proper kernel function to measure the "distance" between two data points.
  - Common kernel functions, like the Gaussian Kernel Function, may be misleading because it ignores the distribution of the data points. We need a kernel function that can reflect the distribution of the data points throughout the data space.
  - This paper introduce Fisher Kernel Method.

Kernel Functions

- By Mercey's Theorem, all kernel function can be written in the following form:

  $K(x_i, x_j) = \Phi_{x_i}{}^T \bullet \Phi_{x_j}$,

  where Φ_{x_i} and Φ_{x_j} are feature vectors of x_i and x_j
- This paper says we can use so called "Fisher Score" to map a vector into its feature vector counterpart.

Fisher Kernel Function

- Define the natural mapping as

  $\Phi_x = I^{-1} \cdot U_x$,   where U_x is the Fisher Score of vector x
- Then our new kernel function will be

$$
\begin{aligned}
K(x_i, x_j) &= \Phi_{x_i}{}^T \cdot \Phi_{x_j} \\
&= (I^{-1} \cdot U_{x_i})^T \cdot (I^{-1} \cdot U_{x_j}) \\
&= U_{x_i}{}^T \cdot (I^{-1})^T \cdot I^{-1} \cdot U_{x_j} \\
&= U_{x_i}{}^T \cdot I^{-1} \cdot I^{-1} \cdot U_{x_j} \quad (\text{by } (I^{-1})^T = I^{-1}) \\
&= U_{x_i}{}^T \cdot I^{-1} \cdot U_{x_j} \qquad (\text{by } (I^{-1})^2 = I^{-1})
\end{aligned}
$$

Exploiting Generative Models in Discriminative Classifiers

- Tommi S. Jaakkola and David Haussler

In Advances in Neural Information Processing Systems, volume 11, 1998.

Mohammad Saif

Professor: R. Maclin

Date : April 2, 2003

Generative Models

- Shakespeares Play

  You too …..

- Automatic generation
  - Given an alphabet

  { you, too, I, thou, kill, happy, Brutus, Ceaser }

  - P(You / You too) … probably not much
  - P(Brutus / You too) … probably a lot more
  - Posterior probability

Classification

- A more practical example
- Classifying DNA gene sequences
  - Consider the alphabet

    { A, B, C, D, E}

  - Classes X and Y
  - Example sequences

    ACDB **EDBC**

  - P(X / ACDB) compared to P(Y / ACDB)

Advantages

- Missing information
  - For example counting methods need counts of an example to deduce probabilities

    Julias Ceaser in anguish said you too Brutus

    - Actual occurrence may not have occurred
- Variable length sequences
  - Utilizes the sequence for classification
  - Not just a bag of words

Discriminative methods

- Often have better performance
- A hybrid of the two methods ideal
- If examples L and M belong to different classes
  - Use difference in generative process
  - Rather than simply the posterior probabilities
- Use probabilities to map the examples into space via a kernel function
- Use discriminative method to classify