MDI341 Structured Data

Energy-based approaches via multi-class classification

Florence d'Alché

Contact: florence.dalche@telecom-paris.fr,
Télécom Paris, Institut Polytechnique de France

From multi class classification to structured prediction

- Scoring functions:
    - Model of the form: $h(x) = \arg\max_{y \in \mathcal{Y}} score(x, y)$
- (1) Solve the problem for multiple classes
- (2) Solve the problem in general for any structured prediction problem (next session)

Document classification

Example 1

- INPUT : ". . . run a health care insurance program ..."
- OUTPUT : politics

Example 2

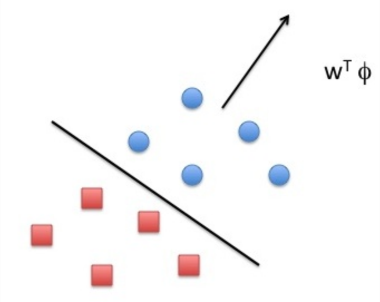
- INPUT: ". . . run the marathon ..."
- OUTPUT : sports

A binary classification task

Two classes: politics, sports

Using a linear model

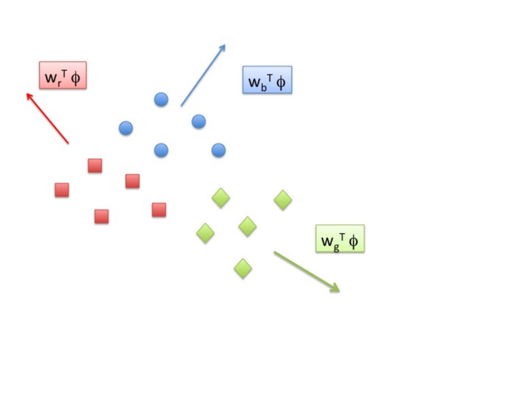
- Input features: for instance, bag of words
- Prediction: $h_w(x) = sgn(w^T \phi(x))$



Now a multiclass classification task

Multiple classes : economics, politics, sports

- Each class y defined by a linear model of the following form: $h_y(x) = w_y^T \phi(x)$



Joint Feature Map for Multiclass classification

With p classes:

$$\begin{aligned} \phi(x,y)^T &= [0\dots0\ \phi(x)^T\ 0\dots0] \\ w^T &= [\mathbf{w}_1^T\dots,\mathbf{w}_y^T,\dots\mathbf{w}_p^T]g \end{aligned}$$

Remember : **here** y is a class label

$$g(x_i,y,w) = \mathbf{w}^T\phi(x_i,y) = \mathbf{w}_y^T\phi(x_i)$$

- Whatever y, w_y's have the same dimension, say p.
- The vector $\mathbf{w}$ is the stack of all $\mathbf{w}_y$ with $y \in$ the finite set $\mathcal{Y}$
- NB : we will note: $\phi(x_i,y) = \phi_i(y)$

Linear Models for Multiclass Classification using Joint Feature Maps

Scoring methods for multiclass classification

$$g(x_i, y, \mathbf{w}) = \mathbf{w}^T \phi(x_i, y) = \mathbf{w}^T \phi_i(y)$$

$$prediction(x_i, \mathbf{w}) = \arg\max_{y \in \mathcal{Y}} \mathbf{w}^T \phi_i(y)$$

Learning Methods

- Structured Perceptron (Collins , EMNLP 2002)
- Logistic regression, CRF (Collins, EMNLP 2002)
- Struct-SVM : Crammer and Singer 2001, Tsochantaridis et al. 2005

1 - Learning linear models: the perceptron rule

Simple discriminative method

$$y' = \arg\max_{y} \mathbf{w}^T \phi_i(y) \tag{1}$$

If $y' \neq y_i$ then, $\mathbf{w} \leftarrow \mathbf{w} + \eta(\phi_i(y_i) - \phi_i(y'))$

Remember the idea of perceptron: if there is a mistake, I add the right vector and substract the wrong vector. If no mistake , I do nothing.
Note that later we will use the following notation:
$\delta\phi_i(y') = (\phi_i(y_i) - \phi_i(y'))$
Collins, 2002.

Learning linear models by minimizing a loss function

- What is a training error here ?
- $\textit{error} = \sum_i \textit{step}(\mathbf{w}^T \phi_i(y_i) - \max_{y \neq y_i} \mathbf{w}^T \phi_i(y))$
- with $\textit{step}(z) = 1$ if $z < 0$ and 0, otherwise
  - zero-one loss : discontinuous, minimization is NP-complete
  - Turn to convexified losses

2 - Log loss, logistic loss

- Posterior probabilities

$$P(y|x, \mathbf{w}) = \frac{\exp(\mathbf{w}^T \phi(x, y))}{\sum_{y'} \exp(\mathbf{w}^T \phi(x, y'))}$$

- Maximize the log conditional likelihood of training data

$$\max_{\mathbf{w}} \log \prod_i P(y_i|x_i, \mathbf{w}) = \sum_i \log \left(\frac{\exp(\mathbf{w}^T \phi_i(y_i)}{\sum_y \exp(\mathbf{w}^T \phi_i(y))} \right)$$

$$\max_{\mathbf{w}} \sum_i (\mathbf{w}^T \phi_i(y_i) - \log \sum_y \exp(\mathbf{w}^T \phi_i(y)))$$

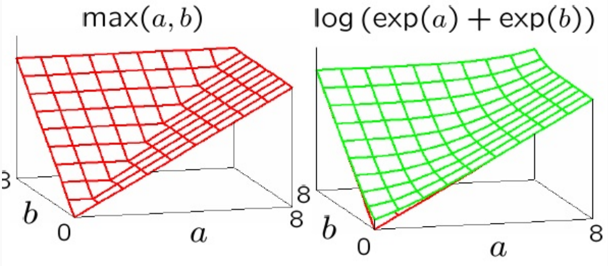
Maximize log loss with regularization

$$\max_{\mathbf{w}} \sum_i (\mathbf{w}^T \phi_i(y_i) - \log \sum_y \exp(\mathbf{w}^T \phi_i(y))) - \lambda ||\mathbf{w}||^2$$

equivalent to

$$\min_{\mathbf{w}} \lambda ||\mathbf{w}||^2 - \sum_i (\mathbf{w}^T \phi_i(y_i) - \log \sum_y \exp(\mathbf{w}^T \phi_i(y)))$$

soft-max



Let us notice the proximity between these two functions.

3 - Now let us try to maximize a margin

If we just want to separate the data we would impose:

$$\forall i, \forall y \neq y_i, \mathbf{w}^T \phi_i(y_i) \geq \mathbf{w}^T \phi_i(y)$$

but we define what is a good separator using the idea of geometric margin !

Now maximizing a margin

On our example:

$\mathbf{w}^T\phi(\textit{run the marathon}, \textit{sports}) \geq \mathbf{w}^T\phi(\textit{run the marathon}, \textit{politics}) + \gamma$

$\mathbf{w}^T\phi(\textit{run the marathon}, \textit{sports}) \geq \mathbf{w}^T\phi(\textit{run the marathon}, \textit{economics}) + \gamma$

$\mathbf{w}^T\phi(\textit{run the marathon}, \textit{sports}) \geq \mathbf{w}^T\phi(\textit{run the marathon}, \textit{sports})$

Let us take $\gamma = 1$ and $\Delta_i(y) = \Delta(y_i, y) = 0$ if $y = y_i$ and $\Delta(y_i, y) = \gamma = 1$, otherwise.

Here and there, $\Delta(y_i, y)$ measures how much y is far from the true output.

Minimizing the norm of "canonical hyperplane

$\min_{\mathbf{w}} \frac{1}{2} \|\mathbf{w}\|^2$

s.t. :

$\forall i, \forall y, \mathbf{w}^T\phi_i(y_i) \geq \mathbf{w}^T\phi_i(y) + \Delta(y_i, y)$

Allowing for non-separability (adding slack variables)

Margin maximization with slack variables: Pb 1

$\min_{\mathbf{w}} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_i \xi_i$

s.t. :

$\forall i, \forall y, \mathbf{w}^T \phi_i(y_i) + \xi_i \geq \mathbf{w}^T \phi_i(y) + \Delta(y_i, y)$

$\forall i, \xi_i \geq 0$

A min-max formulation (Tsochantaridis et al. 2005)

We solve ξ_i: $\forall i, \forall y, \xi_i \geq \mathbf{w}^T\phi_i(y) + \Delta(y_i, y) - \mathbf{w}^T\phi_i(y_i)$
$\forall i, \xi_i = \max_y[\mathbf{w}^T\phi_i(y) + \Delta(y_i, y)] - \mathbf{w}^T\phi_i(y_i)$

Pb 2

$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \max(0, \max_y[\mathbf{w}^T\phi_i(y) + \Delta(y_i, y)] - \mathbf{w}^T\phi_i(y_i))$

Compare max-margin and maxent (log-loss)

Maxent (in logistic regression)

$$\min_{\mathbf{w}} \lambda \|w\|^2 - \sum_i (\mathbf{w}^T \phi_i(y_i) - \log(\sum_i y \exp(\mathbf{w}^T \phi_i(y)))$$

SVM

$$\min_{\mathbf{w}} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_i \max(0, \max_y [\mathbf{w}^T \phi_i(y) + \Delta(y_i, y)] - \mathbf{w}^T \phi_i(y_i))$$

Both try to make the true score better than a function of the other score

What have we seen so far ?

- a simple way to define joint feature map (which will be used as well in structured prediction)
- the structured hinge loss:
  $\ell(y_i, \mathbf{w}^T\phi_i(y)) = \max(0, \max_y[\mathbf{w}^T\phi_i(y) + \Delta(y_i, y)] - \mathbf{w}^T\phi_i(y_i))$
- its proximity with maxent in a logistic regression model

Interestingly the hinge loss allows to take into account the loss Δ between classes y

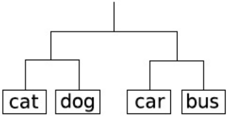
Structured output SVM for a hierarchy of classes

Hierarchical Multiclass Loss:

$$\Delta(y, y') := \frac{1}{2}(\text{distance in tree})$$

$$\Delta(\text{cat}, \text{cat}) = 0, \quad \Delta(\text{cat}, \text{dog}) = 1,$$

$$\Delta(\text{cat}, \text{bus}) = 2, \quad etc.$$



Solve: $$\min_{w,\xi} \frac{1}{2}\|w\|^2 + \frac{C}{N}\sum_{n=1}^{N}\xi^n$$

subject to, for $i = 1, \ldots, n$,

$$\langle w, \phi(x^n, y^n)\rangle - \langle w, \phi(x^n, y)\rangle \geq \Delta(y^n, y) - \xi^n \quad \text{for all } y \in \mathcal{Y}.$$

Tasks solved with this approach

Tsochantaridis, Joachims, Hofman and Altun. JMLR 2005.
This approach for multiple classes can be extended to other kinds of structure.

- multi-class classification
- hierarchical/structured classification
- sequence labelling

References for this lecture

- Crammer, Koby and Singer, Yoram, On the Algorithmic Implementation of Multiclass Kernel-based Vector Machines, J. Mach. Learn. Res., 3/1/2002.
- Tsochantaridis, I. and Joachims, T. and Hofmann, T. and Altun, Y., Large margin methods for structured and interdependent output variables, JMLR, 6,2005
- Collins, Michael, Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms, Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10,2002.
- Ben Taskar, Learning structured prediction models, a large margin approach, PhD thesis (`http://www.seas.upenn.edu/~taskar/pubs/thesis.pdf`), U. Pennsylvany, USA, 2004.