Short course

A vademecum of statistical pattern recognition and machine learning

Conditional random fields

Massimo Piccardi
University of Technology, Sydney, Australia

Agenda

- Exponential family of distributions
- Example: the Gaussian distribution
- Example: the discrete distribution
- Example: a distribution conditioned on a discrete variable
- Graphical models
- Example: the mixture distribution
- Conditional model: conditional random fields
- Learning, maximum conditional likelihood
- Learning with hidden variables
- The linear-chain CRF and the HCRF

Exponential family of distributions

- A distribution is said to belong to the **exponential family** if its pdf can be written as:

$$p(x) = \frac{e^{\langle \theta, f(x) \rangle}}{Z(\theta) = \int_x e^{\langle \theta, f(x) \rangle} dx < +\infty}$$

- The exponential function guarantees that p(x) ≥ 0, as required
- The denominator is a normalization constant (it does not depend on the random variable x) known as *partition function*, > 0 at its turn, that guarantees that ∫p(x) dx = 1. To this aim, Z(θ) must not be infinite (NB: the notation is a bit sloppy as we use x for both the argument of p(x) and the integration variable, all through these slides). Of course, if x is a discrete r.v. the integral is replaced by a sum

Exponential family: notations

$$p(x) = \frac{e^{\langle \theta, f(x) \rangle}}{Z(\theta) = \int_x e^{\langle \theta, f(x) \rangle} dx < +\infty}$$

- < > is the dot, or scalar, product
- **θ** is a vector of parameters, known as *canonical* (aka exponential, natural) *parameters*
- **f(x)** is a vector of functions of the random variable, known as *sufficient statistics* (aka potential functions, feature functions), of the same size as θ <u>and not containing θ</u>

Exponential family: notations

- Please note that all the following notations for the scalar product are equivalent!

$$\langle \theta, f(x) \rangle = \langle f(x), \theta \rangle =$$
$$= \theta^T f(x) = f(x)^T \theta$$
$$\theta \cdot f(x) = f(x) \cdot \theta$$
$$= \sum_{k=1}^{K} \theta_k f_k(x)$$

Exponential family: notations

Do not be confused:

- given **G(θ) = 1/Z(θ)**, some authors prefer to write:

$$p(x) = G(\theta)\, e^{\langle \theta, f(x) \rangle}$$

- given **A(θ) = ln Z(θ)** (known as *log-partition function* or *cumulant function*), other authors prefer to write:

$$p(x) = \frac{e^{\langle \theta, f(x) \rangle}}{Z(\theta)} = e^{\langle \theta, f(x) \rangle} e^{\ln\left(\frac{1}{Z(\theta)}\right)} = e^{\langle \theta, f(x) \rangle - A(\theta)}$$

Exponential family: notations

- Others (e.g., Bishop, Wikipedia) prefer to separate a term in x useful to later show certain properties of conjugacy:

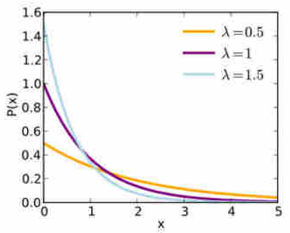
$$p(x) = H(x)G(\theta)\, e^{\langle \theta, f(x) \rangle}$$

- However, this notation does not expand the distributions covered:

$$p(x) = H(x)G(\theta)\, e^{\langle \theta, f(x) \rangle} = e^{\ln H(x)} G(\theta)\, e^{\langle \theta, f(x) \rangle} =$$
$$= G(\theta)\, e^{\langle \theta, f(x) \rangle + 1 \cdot \ln H(x)} = G(\theta)\, e^{\langle \theta', f'(x) \rangle}$$

Exponential family: examples

- Members: Gaussian, exponential, gamma, chi-squared, beta, Dirichlet, discrete/categorical, Bernoulli, binomial, multinomial, Poisson, Wishart, Inverse Wishart and many others (Wikipedia, Nov. 2011)
- Notable non-members: Cauchy, Student's t, Laplace (for mean $\neq 0$)



The exponential distribution, $p(x) = \lambda e^{-\lambda x}$ (courtesy of Wikipedia)

Example: the Gaussian distribution

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{\mu^2}{2\sigma^2}} =$$

$$= e^{\ln\frac{1}{\sqrt{2\pi\sigma^2}}} e^{-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{\mu^2}{2\sigma^2}} = e^{-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \left(\frac{\mu^2}{2\sigma^2} + \frac{1}{2}\ln\left(2\pi\sigma^2\right)\right)} =$$

$$= e^{\theta_2 x^2 + \theta_1 x - A(\theta_1, \theta_2)}$$

$$\rightarrow \begin{array}{l} f_1(x) = x,\ f_2(x) = x^2 \\ \theta_1 = \frac{\mu}{\sigma^2},\ \theta_2 = -\frac{1}{2\sigma^2} \end{array}$$

Canonical and mean parameters

- Canonical parameters θ_1, θ_2 (alongside feature functions f(x)) offer a full parametrisation for the Gaussian distribution. This is an alternative to the usual μ, σ^2 (or $E[x^2] = \sigma^2 + \mu^2$) parameters which are known as the *mean parameters*

- Mapping from canonical to mean for the Gaussian:

$$\mu = -\frac{\theta_1}{2\theta_2},\ \sigma^2 = -\frac{1}{2\theta_2}$$

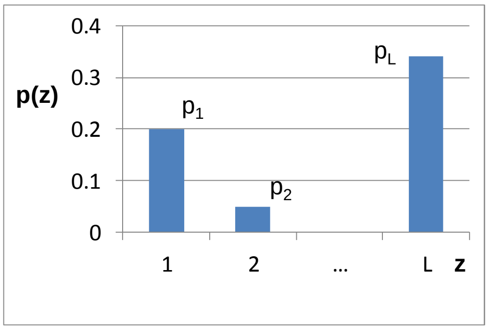
Example: the discrete distribution

- A discrete distribution (or categorical) over a discrete random variable, z, with L possible outcomes is defined by L probability values, p_l, l = 1...L (one is redundant)
- $z \in \{1,...l,...L\}$ $\quad 0 < p_l \leq 1$, l = 1...L $\quad \Sigma_l\, p_l = 1$
- Various notations can be used to represent the above compactly. Examples:

$$p(z) = \sum_{l=1}^{L} p_l \cdot \mathrm{I}(z = l) = \prod_{l=1}^{L} p_l^{\mathrm{I}(z=l)}$$

- z may also be expressed in 1-out-of-N notation: $[z_1=0/1, z_2=0/1,.., z_L=0/1]$ with changes to the probability notation

The discrete distribution



$$p(z) = \prod_{l=1}^{L} p_l^{\mathrm{I}(z=l)}$$

Example: the discrete distribution

- The discrete distribution can be placed in exponential form as:

$$p(z)=\prod_{l=1}^{L} p_l^{\mathrm{I}(z=l)}=\prod_{l=1}^{L}\left(e^{\ln p_l}\right)^{\mathrm{I}(z=l)}=$$

$$=\prod_{l=1}^{L} e^{\ln p_l \cdot \mathrm{I}(z=l)}=e^{\sum_{l=1}^{L}\ln p_l \cdot \mathrm{I}(z=l)}=e^{\sum_{l=1}^{L}\theta_l \cdot \mathrm{I}(z=l)}=e^{\langle \theta, f(z)\rangle}$$

$$\begin{aligned} &\theta_l=\ln p_l \\ \rightarrow\ &f_l(z)=\mathrm{I}(z=l) \\ &A(\theta)=0 \end{aligned}$$

Example: a distribution conditioned on a discrete variable

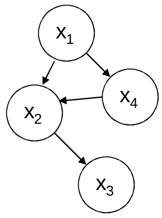
- As further example, imagine having a Gaussian distribution <u>per class</u>, like in class-conditional likelihoods
- We can note this case as p(x|z), where x is the Gaussian random variable and z the class index
- The exponential notation could be:

$$p(x\,|\,z)=\frac{e^{\sum_{l=1}^{L}\langle \theta_l, f(x)\rangle \mathrm{I}(z=l)}}{Z(\theta)}$$

Graphical models

One random variable in each node of a graph

direct acyclic graph (DAG)



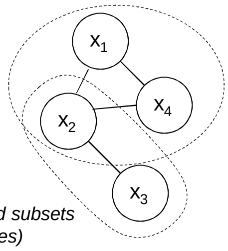
- Directed graphical models (DGM)

$$p(x_1, \ldots, x_M) = \prod_{i=1}^{M} p\left(x_i \mid x_{parents(i)}\right)$$

- Undirected graphical models (UGM)

$$p(x_1, \ldots, x_M) = \frac{1}{Z} \prod_{c=1}^{C} \psi_{subset\ c}\left(x_{subset\ c}\right)$$

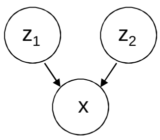
potential functions, $\in \Re+$



fully connected subsets (cliques)

Factorisation

- With directed graphs, we need to work with normalised probabilities and factorise according to Bayes' theorem and independence; undirected graphs are less constrained
- Perfect correspondence is not always possible. Example: given three arbitrary variables x, z_1, z_2, this directed model:

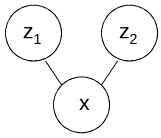


factorises as $p(x, z_1, z_2) = p(x \mid z_1, z_2)\, p(z_1)\, p(z_2)$

If z_1 and z_2 are discrete with L values each, $p(z_1)$ and $p(z_2)$ have a total of 2(N – 1) dof and there are N^2 different $p(x \mid z_1, z_2)$

Factorisation

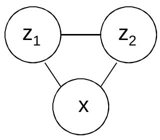
- Which undirected graph corresponds to the previous directed graph? If we consider this graph:



$$p(x, z_1, z_2) = \frac{1}{Z} \psi_1(x, z_1) \psi_2(x, z_2)$$

we cannot represent the joint dependence of x from z_1, z_2

- If we consider this graph (the *moral graph*, Bishop 8.3.4), we lose the independence between z_1 and z_2



$$p(x, z_1, z_2) = \frac{1}{Z} \psi_1(x, z_1, z_2)$$

UGM and exponential family

- Given that $\psi > 0$, we can also express the UGM with exponential functions:

$$p(x_1, \dots, x_M) = \frac{1}{Z} \prod_{c=1}^{C} \psi_{subset\ c}\left(x_{subset\ c}\right) = \frac{1}{Z} \prod_{c=1}^{C} e^{\ln \psi_{subset\ c}\left(x_{subset\ c}\right)} =$$

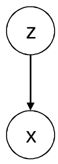
$$= \frac{1}{Z} \prod_{c=1}^{C} e^{U_{subset\ c}\left(x_{subset\ c}\right)}$$

- We assume the **log-linear model** for U, making the UGM equivalent to the exponential family:

$$U\left(x_1 \dots x_{C_M}\right) = \left\langle \theta, f\left(x_1 \dots x_{C_M}\right) \right\rangle = \sum_{k=1}^{K} \theta_k f_k\left(x_1 \dots x_{C_M}\right)$$

Example: the mixture distribution

- Consider a mixture distribution with L components where z is the discrete component indicator and x is a continuous measurement. We write the generative model as:



$$p(x,z) = p(x \mid z)\,p(z)$$

- We have shown that p(z) is a member of the exponential family: if also p(x|z) is a member, we can write the joint probability, p(x,z), conditional probability p(z|x) and marginal probability p(x) in terms of the exponential family

Example: the mixture distribution

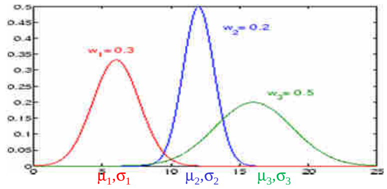
- The joint probability is:

$$p(x,z) = \frac{e^{\sum_{l=1}^{L}\langle \theta_l^{[x]}, f(x)\rangle \mathrm{I}(z=l)}\; e^{\sum_{l=1}^{L}\theta_l^{[z]}\mathrm{I}(z=l)}}{Z(\theta)} = \frac{e^{\sum_{l=1}^{L}\left(\langle \theta_l^{[x]}, f(x)\rangle + \theta_l^{[z]}\right)\mathrm{I}(z=l)}}{\int_x \sum_z e^{\sum_{l=1}^{L}\left(\langle \theta_l^{[x]} f(x)\rangle + \theta_l^{[z]}\right)\mathrm{I}(z=l)}\,dx}$$

where $\theta_l^{[x]}$ and $\theta_l^{[z]}$ are the canonical parameters for p(x|z=l) and p(z=l), respectively. $\theta_l^{[x]}$ is a vector of adequate size, while $\theta_l^{[z]}$ is a scalar

- NB: we can also compound the two terms $\theta_l^{[x]}$ and $\theta_l^{[z]}$ into a single θ_l by adding a 1 to the feature vector: $f'(x)^T = [f(x)\ 1]^T$

Example: GMM



with the usual mean parameters:

$$p(x, z = l) = w_l \, N\left(x \mid \mu_l, \sigma_l^2\right)$$

with the canonical parameters:

$$p(x, z = l) = \frac{e^{\left\langle \theta_l^{[x]}, f(x) \right\rangle + \theta_l^{[z]}}}{Z}$$

- Vector $\theta_l^{[x]}$ corresponds to parameters $\{\mu_l, \sigma_l^2\}$
- $\theta_l^{[z]}$ corresponds to parameter w_l
- Their sum in log scale is equivalent to a product in linear scale

Mixture distribution: marginal for x

- With the compacted notation, the joint probability is:

$$p(x, z) == \frac{e^{\sum_{l=1}^{L} \langle \theta_l, f'(x) \rangle \mathrm{I}(z=l)}}{\int_x \sum_z e^{\sum_{l=1}^{L} \langle \theta_l, f'(x) \rangle \mathrm{I}(z=l)} dx}$$

Like with the mean parameters, marginal p(x) requires an explicit summation:

$$p(x) = \sum_{l=1}^{L} p(x, z = l) == \frac{\sum_{l=1}^{L} e^{\sum_{l=1}^{L} \langle \theta_l, f'(x) \rangle \mathrm{I}(z=l)}}{\int_x \sum_z e^{\sum_{l=1}^{L} \langle \theta_l, f'(x) \rangle \mathrm{I}(z=l)} dx}$$

NB: p(x) is not a member of the exponential family!

Inference: the conditional model

- The conditional probability of z given x can be expressed as:

$$p(z \mid x) = \frac{p(x,z)}{p(x)} = \frac{p(x \mid z)p(z)}{\sum_z p(x \mid z)p(z)}$$

- With the previous hypotheses:

$$p(z \mid x) = \frac{e^{\sum_{l=1}^{L}\langle \theta_l, f'(x)\rangle \mathrm{I}(z=l)}}{\int_x \sum_z e^{\sum_{l=1}^{L}\langle \theta_l, f'(x)\rangle \mathrm{I}(z=l)}dx} \Bigg/ \frac{\sum_z e^{\sum_{l=1}^{L}\langle \theta_l, f'(x)\rangle \mathrm{I}(z=l)}}{\int_x \sum_z e^{\sum_{l=1}^{L}\langle \theta_l, f'(x)\rangle \mathrm{I}(z=l)}dx} =$$

$$= e^{\sum_{l=1}^{L}\langle \theta_l, f'(x)\rangle \mathrm{I}(z=l)} \Big/ \sum_z e^{\sum_{l=1}^{L}\langle \theta_l, f'(x)\rangle \mathrm{I}(z=l)}$$

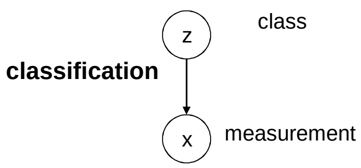
***conditional random field**:*
a conditional distribution in correspondence with an undirected graphical model

NB: the conditional model is simpler than the joint model!

 23

The conditional model: classification

- The graphical model of classification is identical to that of a mixture distribution!:



One can look at z as a class label and x as its measurement; the conditional model, p(z|x), is therefore the *inference* of z given x

- Given x and p(z|x), apply a decision rule such as MAP or minimum expected risk to obtain the class
- This classifier is called the **logistic regression** classifier

 24

The logistic regression classifier

$$p(z \mid x) = \frac{e^{\sum_{l=1}^{L} \langle \theta_l , f'(x) \rangle \mathrm{I}(z=l)}}{\sum_z e^{\sum_{l=1}^{L} \langle \theta_l , f'(x) \rangle \mathrm{I}(z=l)}}$$

- z is the class and f'(x) the measurement, or a fixed manipulation of the measurement. The logistic regression classifier is the simplest conditional random field

The logistic regression classifier

- With even simpler notations:

$$p(z \mid x) = \frac{e^{\langle \theta_z , f'(x) \rangle}}{\sum_z e^{\langle \theta_z , f'(x) \rangle}}$$

- The exponent is just a scalar product between the parameters of class z, θ_z, and the feature functions, f'(x)

MAP classification with the logistic regression classifier

- Please note:

$$z^* = \arg\max_z \, p(z \mid x) =$$
$$= \arg\max_z \, e^{\langle \theta_z , f'(x) \rangle} =$$
$$= \arg\max_z \, \langle \theta_z , f'(x) \rangle \quad !$$

- First, the denominator does not depend on z and therefore does not count in the decision. Second, the logarithm of the numerator has the same maximum as its argument. With the MAP rule, this is just a linear classifier!

Learning

- A model can be learned from samples with maximum likelihood
- *Unsupervised learning:* we assume that the samples are only the N measurements, $x_{1:N}$, without knowledge of the labels
- The likelihood function is:

$$p(x_{1:N} \mid \theta) = \prod_{n=1}^{N} p(x_n \mid \theta) = \prod_{n=1}^{N} \sum_{z_n=1}^{L} p(x_n, z_n \mid \theta) = \prod_{n=1}^{N} \sum_{z_n=1}^{L} p\left(x_n \mid z_n, \theta^{[x]}\right) p\left(z_n \mid \theta^{[z]}\right)$$

- MLE finds parameters θ maximising the above, where we have marginalised the z_n. As you know, this function is not concave/log-concave (has multiple local maxima) and EM is a classic solver

Learning

- Let us now assume that we know the N measurements, $x_{1:N}$, and their corresponding labels, $z_{1:N}$.This is *supervised learning* and is a common assumption in classification
- We can now write the (joint) likelihood function as:

$$p(x_{1:N}, z_{1:N} \mid \theta) = \prod_{n=1}^{N} p(x_n, z_n \mid \theta) = \prod_{n=1}^{N} p(x_n \mid z_n, \theta^{[x]}) p(z_n \mid \theta^{[z]})$$

- The objective function for the MLE is now much simpler since we assume knowledge of the z_n. The maximisation for $\theta^{[z]}$ certainly has a unique, closed-form solution. The maximisation for $\theta^{[x]}$ has a unique solution depending on the choice for p(x|z). For instance, if p(x|z) belongs to the exponential family, the solution is unique

Learning

- If we have a conditional model available, we can also write the **conditional likelihood function** as:

$$p(z_{1:N} \mid x_{1:N}, \theta) = \prod_{n=1}^{N} p(z_n \mid x_n, \theta)$$

- If at “run time” we are interested in solving the inference, p(z|x), the above MCLE is often very effective since it explicitly trains the inference model itself (**discriminative training**)
- The MCLE of exponential family is a convex problem in θ!
- All these MLE can be turned into corresponding MAPE by adding a prior on θ (this includes regularisers such as L2 and L1 norms)

Learning

- The conditional likelihood of exponential family is a concave function in θ. Example with supervised classification:

$$\ln p(z_{1:N} \mid x_{1:N}, \theta) = \sum_{n=1}^{N} \ln p(z_n \mid x_n, \theta) =$$

$$= \sum_{n=1}^{N} \ln \left(e^{\sum_{l=1}^{L} \langle \theta_l, f'(x_n) \rangle \mathrm{I}(z_n = l)} \Big/ \sum_{z} e^{\sum_{l=1}^{L} \langle \theta_l, f'(x_n) \rangle \mathrm{I}(z = l)} \right) =$$

$$= \sum_{n=1}^{N} \left(\sum_{l=1}^{L} \langle \theta_l, f'(x_n) \rangle \mathrm{I}(z_n = l) - \ln \sum_{z} e^{\sum_{l=1}^{L} \langle \theta_l, f'(x_n) \rangle \mathrm{I}(z = l)} \right)$$

- The above follows from the fact that a function of the form log-sum-exp (the 2nd term) is convex. The 1st term is linear (both convex or concave)
- The global maximum for the conditional likelihood can be found with numerical algorithms such as quasi-Newton methods (L-BFGS)

Learning with hidden variables

- If the conditional model also contains hidden variables (variables which are not observed at training time), the MCLE is not convex anymore!
- Example:

$$p(z \mid x) = \sum_{h} p(z, h \mid x) = \frac{\sum_{h} e^{\cdots}}{\sum_{z} \sum_{h} e^{\cdots}}$$

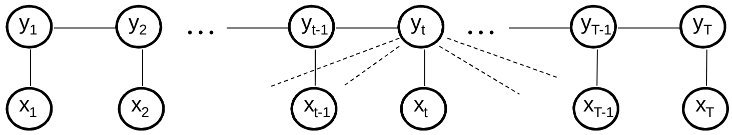
NB: p(z|x) is not a member of the exponential family (sum-exp)

$$\sum_{n=1}^{N} \ln p(z_n \mid x_n, \theta) = \sum_{n=1}^{N} \ln \sum_{h} p(z_n, h_n \mid x_n) =$$

$$= \sum_{n=1}^{N} \ln \sum_{h} \exp(\ldots) - \sum_{n=1}^{N} \ln \sum_{z} \sum_{h} \exp(\ldots)$$

- Convex + concave is neither convex nor concave [see, for instance, S. Gong, T. Xiang, Visual Analysis of Behaviour: From Pixels to Semantics, 2011]
- Solvers: L-BFGS, Generalised EM, Stochastic Gradient Descent (SGD)

The linear-chain CRF

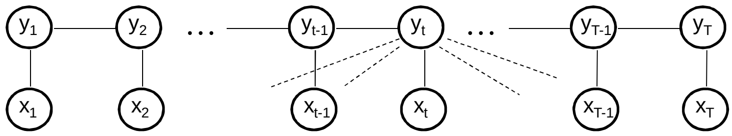


- Same graphical model as the HMM. The conditional model is:

$$p(y_{1:T} \mid x_{1:T}, \theta)$$

- The feature functions may contain other observations than just x_t
- Learning is supervised! (states are assumed known)
- Decoding ($y^*_{1:T} = \text{argmax } p(y_{1:T}|x_{1:T},\theta)$): Viterbi-like

The linear-chain CRF

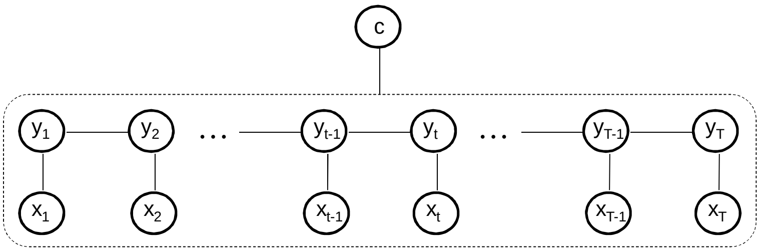


$$p(y_{1:T} \mid x_{1:T}) = \frac{e^{\sum_{l=1}^{L} \langle \theta_l, f(x_1) \rangle \mathrm{I}(y_1 = l)} \; e^{\sum_{l=1}^{L} \sum_{k=1}^{L} \theta_{lk} \mathrm{I}(y_1 = l, y_2 = k)} \cdots e^{\sum_{l=1}^{L} \langle \theta_l, f(x_T) \rangle \mathrm{I}(y_T = l)}}{\sum_{y_1 \cdots y_T} (\cdots)}$$

$$\propto p(x_t \mid y_t)$$

$$\propto p(y_t \mid y_{t-1})$$

The hidden (linear-chain) CRF (HCRF)



- Analogous to one HMM per class c. The conditional model is:

$$p(c \mid x_{1:T}, \theta) = \sum_{y_{1:T}} p(c, y_{1:T} \mid x_{1:T}, \theta)$$

- Decoding ($c^* = \arg\max p(c|x_{1:T},\theta)$): forward formula-like
- Learning: hidden variables case (states are unsupervised in this case)

Software

- HCRF library (Louis-Philippe Morency, version HCRF2.0b, March 2011; C++, Matlab, Python)
  - models and solvers for CRF, HCRF, LDCRF (latent dynamic CRF)
- UGM: Matlab code for undirected graphical models (Mark Schmidt, 2011 version)
- Conditional Random Field (CRF) Toolbox for Matlab (Kevin Murphy), probably now incorporated into
- PMTK3 (probabilistic modeling toolkit for Matlab/Octave, version 3, Kevin Murphy)
- others

References

- J. D. Lafferty, A. McCallum, F. C. N. Pereira, “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data,” ICML 2001: 282-289
- Sutton, C., McCallum, A.: An Introduction to Conditional Random Fields for Relational Learning. In "Introduction to Statistical Relational Learning," Lise Getoor and Ben Taskar eds, MIT Press (2006)
- R. Klinger, K. Tomanek, “Classical Probabilistic Models and Conditional Random Fields,” TR07-2-013, TU Dortmund, 2007
- Douglas L. Vail, Manuela M. Veloso, John D. Lafferty, “Conditional random fields for activity recognition,” AAMAS 2007: 235, 8 pages
- Sy Bor Wang, Quattoni, A., Morency, L.-P., Demirdjian, D., Darrell, T., “Hidden Conditional Random Fields for Gesture Recognition" in Proc. CVPR 2006, pp. 1521-1527
- A. Quattoni, Sy Bor Wang, L-P. Morency, M. Collins, T. Darrell, “Hidden Conditional Random Fields,” IEEE Trans. Pattern Anal. Mach. Intell. 29(10): 1848-1852 (2007)
- M. Wainwright, M. Jordan’s book: “Graphical models, exponential families, and variational inference,” Foundations and Trends in Machine Learning, Now Publishers, 2008
- Stephen Boyd, Lieven Vandenberghe, Convex Optimization, Cambridge University Press, 2004, http://www.stanford.edu/~boyd/cvxbook/

