MEMM (Maximum entropy markov networks)

This Paper presents a new Markovian model, Maximum entropy markov model closely related to hidden markov models that allows many observations to be represented as features. This paper starts with an introduction to Hidden markov models, their representation, factors, Inference algorithm (Viterbi) and how it could be extended to MEMM (maximum entropy markov networks) and their applications. HMM's are composed of finite set of states S, set of possible observations O, two probability distribution tables transition probability from s' to s P(s/s') s,s' in S ,and an observation probability distribution P(o/s) o in O , s in S and an initial state P_o(s) distribution. MEMM combines transition and observation functions in HMM and replace them with a single distribution P(s/o,s') that captures the probability of present state s given the previous state s' and current observation o. We can split this into |S| transition functions $P_{s'}$(s/o) =P(s/o,s'). Each of which could be modelled independently. This kind of representation would allow us define many overlapping features, Which we could not do in HMM's. Further it also solves the problem of maximum likelihood in HMM's. HMM parameters to maximize the likelihood of the observation sequence; however, in most text applications, the task is to predict the state sequence given the observation sequence. In some application It would be difficult to enumerate all possible observation and build a distribution table P(o/s) over observation given states.

Next subsequent sections describes the exponential representation of factors, Parameter Estimation techniques to find the best model using the training data using iterative gradient scaling and presented some results. I would like to conclude the section with MEMM standing in a way between Conditional Random Fields and Hidden markov models which offer increased freedom in choosing features to represent observations. One advantage MEMM over CRF is the ability to train efficiently, in CRF's one needs to use some version of forward-backward algorithm as an inner loop in training. In MEMM's estimating parameters of the maximum entropy distributions used for transition probabilities can be used for each transition distribution in isolation.

Conditional Random Fields: Probabilistic Model for Segmenting and labelling sequence data:

This paper presents Conditional Random field framework to segment and label sequence data and its advantages over Hidden markov models and maximum entropy markov networks. This paper starts with applications of labelling sequence data in various fields such as in computational biology, POS tagging, Information extraction. Then this paper presented a couple of points which could not be dealt with HMM's and MEMM models and the requirement of CRF model. One such reasons is probability of transition may not only depend on present observations but may also depend on past and future observation which is not captured in MEMM and HMM. MEMM could solve some of the issues of HMM using exponential model of observation features of input and outputs a distribution over next possible states. But it could not capture long term future dependencies of sequential data. The paper also presents Label bias problem which could not be dealt completely using MEMM's but is solved in CRF model. MEMM uses per state exponential model to decide the next state, Some times if the next state is single it has no choice but to pass all its mass to next state which could not be case. More generally, states with low entropy next states will take little notice of observations.

Couple of solutions to take care of label bias problem in MEMM are also proposed involving determinisation and fully connected model, but they does not take account of prior structural knowledge of the data.

Then he presents the formal definition for conditional random field model and likelihood function to be maximized. He then gives the notation of features to capture arbitrary dependencies on the observation sequence. One advantage we get with this model is it doesn't need much training data as features do not need to specify completely a state or observation. Another important property is the convexity of loss function which could be optimised much easily. It happens because of convexity of exponential functions.

Then he presented couple of algorithms involving iterative scaling to estimate the parameters of the model. He defines forward and backward vectors over the observation data and gives a formula which look similar to gradient of loss function in a neural network.

The paper then presents experiments and results section emphasizing the improvement over MEMM's and HMM's. He then concludes with a statement that CRF is the first model to combine benefits of conditional models with the global normalization of random field model. I would say CRF's perform better than HMM's and MEMM's whenever there are higher order dependencies than the model.