**Conditional Random Fields**

CRF are a variant of Markov Networks. This type of models are intended to deal with Task-specific prediction, where we have a set of input (observed) variables, and a set of target variables x , and we are trying to predict the target variables y. These models are designed for those cases where we always have the same types of variables as an input variable and the same type of variables as the output/target variables. Example: image segmentation, here we have input which is always a set of pixels (pixel values) - x (which we can process to produce more expressive features, like histograms of colors and textures). And the target variable y is a class for every pixel. We design the model so as to solve the problem of going from x to y. We want to solve that problem (going from x to y). So why don’t we use the regular old graphical model for solving this? Let’s see what might be the issue with using the regular graphical models. So if we are trying to predict a label Ci of a specific super pixel i, having the pixel values/features (color and texture features). These features are more informative about the class of the pixel i and are very correlated with each other. The texture histograms, which tell us the directions in which lines go in a particular super pixel, they are very redundant, in terms of the texture that they measure. So if we have very correlated features containing lots of redundant information, if we represent this in a very naïve bayes model, where the features should be independent given the label // P(X1|Y),P(X2|Y),…P(xn|Y) we are ignoring that correlation structure. That means we are pushed towards very skewed probability distributions, that are not really a good reflections of my data (our true beliefs). They make incorrect independence assumption. So let’s make correct independence assumptions. Lets add a bunch of edges (between all these features) that capture the correlations, but this turns out to be really hard – hard to figure out and gives rise to densely connected models.  
So a completely different solution to this problem says: I don’t care about the (image) features, I don’t want to predict the probability distributions of our pixels/features. //I think P(X). What we are trying to do is to use the given features X, and only care about is to model the distribution over Y. //P(Y|X). So we reformulated our problem. In standard modeling, we model the joint distribution P(X,Y) //probability of X and Y together. Instead of doing this, we will model, a conditional distribution of Y given X: P(Y|X), where we are NOT trying to capture the distribution over X (the features). So if we are not trying to capture the distribution over X we don’t care about the correlation of the features. We can do the modeling of P(Y|X) with the Conditional Random Fields – CRF. CRF at first glance looks just as Gibbs distribution, where we had a set of factors (in my case they are pixel wise probabilities), and to get P(X,Y) = P(X1, X2,X3,…Xn,Y) = P(X1|Y)\*P(X2|Y)\*…\*P(Xn|Y). This is where the difference comes in: if I want to model conditional distribution of Y|X // P(Y|X), 09.07 min, we need to set X on the right hand side,   
which means we need to have a separate normalization constant or a partition function: Z(X) = sum\_Y( P(X,Y)), which is a function of X. What does this means? For any given X, we sum over Y, and we re going to construct the conditional distribution Y over X: P(Y|X) = 1/Z(X) \*P(X,Y) by normalizing the P(X,Y) by this Z partition function/constant. To be a little bit more concrete, hear 10.28 min from he video.

CRF are highly related to the Logistic Model/Function, see 14 min from the video. The logistic model is a very simple CRF. So if we were using/doing the standard joint density modeling of P(X,Y), since the assumption there is that the features X1,X2,…Xn are independent, in our standard modeling for P(X,Y) we would have something similar as the naïve bayes, since we only have pairwise terms which relate the Xi with Y: P(X1|Y), P(X2|Y),…P(Xn|Y) and we don’t have any terms that relate X1, X2,…Xn (the features) to each other. And we know that with the Naïve Bayes we have a very strong independence assumption of the X1, X2,..Xn features. And with the conditional distribution modeling: P(Y|X) we remove from the analysis every notion of a correlation between the X1,X2,..Xn features. We are just modeling how the X1,X2,…Xn come together to affect the probability of Y. That’s really the difference between the Naïve Bayes and Logistic model. The same intuition extends to much richer models with more variables, having more values and more classes. So with these CRF models we have the ability to ignore the distribution over the features and focus only on the target variables, and allow us to sort of ignore correlations between features and not worry about whether they are independent of each other or not.   
What is commonly done: train a discriminatice classifier (SVM, boosting, Random Forest, etc.) anything that we like to predict the probability P(Yi|X). And in fact, that’s how we can achieve a high performance on these tasks (the example was image segmentation). Training a strong classifiers for our node potentials – P(Y|X) and adding on top of that pairwise or high order potentials between the Yi’s.

**Summary**

* CRF is like any other Gibbs distribution, but a critical and subtle difference is that is normalized differently. It is normalized so that we are creating a conditional distribution on Y given X.
  + A special/simple CRF is the standard logistic regression model
* We don’t need to model the distribution over variables X1,X2..Xn that we don’t care about, only we model over Y1,Y2…Yn, the ones we care about predicting.
* We can design really impressive predictors without worrying about wrong independencies (wrong independencies assumptions) between different variables (in this case X1, X2… Xn).