MACHINE LEARNING AND PATTERN RECOGNITION

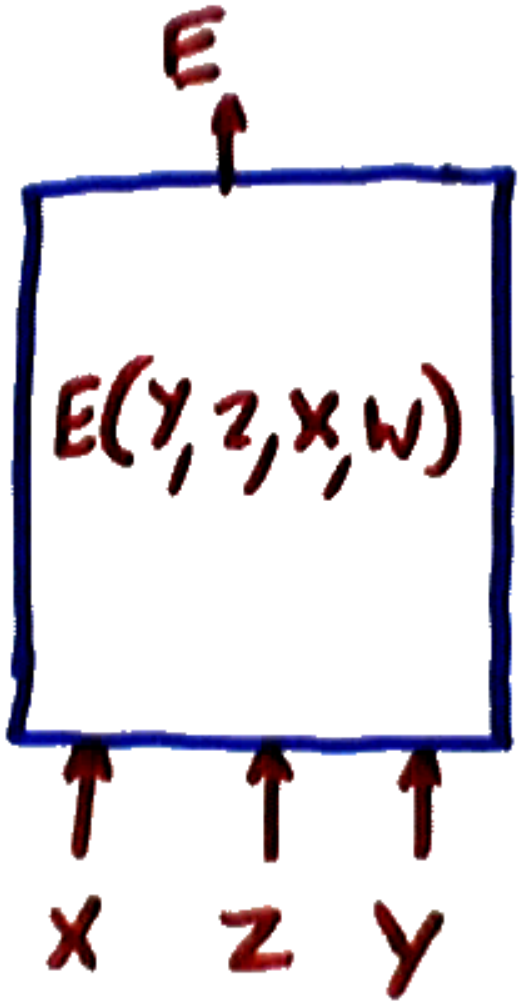
Fall 2006, Lecture 8:

Latent Variables, EM

Yann LeCun
The Courant Institute,
New York University
http://yann.lecun.com

Latent Variables

Latent variables are unobserved random variables Z that enter into the energy function $E(Y, Z, X, W)$.



The X variable (input) is always observed, the Y must be predicted. The Z variable is *latent*: it is not observed. We need to *marginalize* the joint probability $P(Y, Z|X, W)$ over Z to get $P(Y|X, W)$:

$$P(Y|X, W) = \int P(Y, z|X, W)dz$$

The following discussion treats the case where an observation X is present. In the unsupervised case, there is no observation. We can simply remove the symbol X from all the slides below.

Latent Variables: example

Let's say we have a bunch of images of a Boeing 747 under various viewing angles (let's call the angle Z), and another bunch of images of an Airbus A-380, also under various viewing angles.
Let's assume that we are given a "similarity" function $E(Y, Z, X)$ where Y is the label (Boeing or Aribus), Z is the latent variable (the viewing angle), and X the image. For example, $E(\text{Airbus}, 20, X)$ will give us a low energy if X is similar to our prototype image of an Airbus under 20 degree viewing angle. For example, E could be defined as:

$$E(Y, Z, X) = ||X - R_{YZ}||^2$$

where R_{YZ} is our prototype image of plane Y at angle Z.
When asked about the category of an image, we are never given the viewing angle, but knowing it would make our task simpler.

Latent Variables: marginalization

In terms of energy function, $P(Y, Z|X, W)$ can be written as:

$$P(Y, Z|X, W) = \frac{\exp(-\beta E(Y, Z, X, W))}{\int \exp(-\beta E(y, z, X, W)) dz dy}$$

Therefore, $P(Y|X, W) = \int P(Y, z|X, W) dz$ becomes:

$$P(Y|X, W) = \int \frac{\exp(-\beta E(Y, z, X, W))}{\int \exp(-\beta E(y, z', X, W)) dz' dy} dz$$

since the denominator doesn't depend on z:

$$P(Y|X, W) = \frac{\int \exp(-\beta E(Y, z, X, W)) dz}{\int \exp(-\beta E(y, z', X, W)) dz' dy}$$

If Z is a multidimensional variable, this could be very difficult to compute.

Latent Variables: example of marginalization

$$E(Y, Z, X) = ||X - R_{YZ}||^2$$

$$P(Y, Z|X, W) = \frac{\exp(-\beta||X - R_{YZ}||^2)}{\int \exp(-\beta||X - R_{YZ}||^2 dz dy}$$

It's a Gaussian with mean R_{YZ}, and variance $1/\beta$.

$$P(\text{Airbus}|X) = \sum_Z \frac{\exp(-\beta||X - R_{\text{Airbus } Z}||^2)}{\sum_Z \exp(-\beta||X - R_{\text{Boeing } Z}||^2) + \exp(-\beta||X - R_{\text{Airbus } Z}||^2)}$$

It's a sum of Gaussians.

Latent Variables: max likelihood inference

Very often, given an observation X, we merely want to know the value of Y that is the most likely: $Y^* = \text{argmax}_Y P(Y|X,W)$

$$Y^* = \text{argmax}_Y \frac{\int \exp(-\beta E(Y,z,X,W))dz}{\int \exp(-\beta E(y,z',X,W))dz'dy}$$

Since the denominator does not depend on Y, we can simply remove it:

$$Y^* = \text{argmax}_Y \int \exp(-\beta E(Y,z,X,W))dz$$

By taking log and dividing by β, we get:

$$Y^* = \text{argmin}_Y -\frac{1}{\beta}\log\left[\int \exp(-\beta E(Y,z,X,W))dz\right]$$

This is the *logsum* of the energies for all values of Z, also called the *Helmholtz free Energy* of the ensemble of states when Z varies.

Latent Variables: example of max likelihood

$$E(Y, Z, X) = ||X - R_{YZ}||^2$$

$$Y^* = \text{argmin}_Y - \frac{1}{\beta} \log \left[\sum_z \exp(-\beta ||X - R_{YZ}||^2) \right]$$

Latent Variables: zero-temperature limit

Computing the most likely Y using the free energy:

$$Y^* = \text{argmin}_Y - \frac{1}{\beta} \log \left[\int \exp(-\beta E(Y, z, X, W)) dz \right]$$

still requires to compute a (possibly horrible) integral over Z.
One possible shortcut is to make β go to infinity. Then, as we have seen before, the *logsum* reduces to the *min*, hence:

$$\lim_{\beta \to \infty} Y^* = \text{argmin}_Y \min_Z E(Y, Z, X, W)$$

In this case, inference is a lot simpler: to find the "best" value of Y, find the combination of values of both Z and Y that minimize the energy:

$$E(Y^*, Z^*, X, W) = \min_{Y,Z} E(Y, Z, X, W)$$

and return Y^*.

Latent Variables: example of zero-temp limit

$$E(Y, Z, X) = ||X - R_{YZ}||^2$$

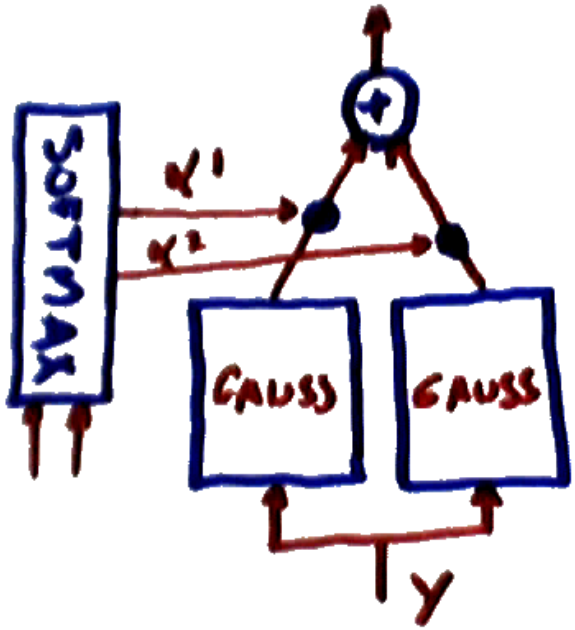
$$E(Y^*, Z^*, X, W) = \min_{Y,Z} ||X - R_{YZ}||^2$$

and return Y^*.

Example: Mixture Models

We have K normalized densities $P^k(Y|W^k)$, each of which has a positive coefficient α^k (whose sum over k is 1), and a switch controlled by a discrete latent variable Z that picks one of the component densities. There is no input X, only an "output" Y (whose distribution is to be modeled) and a latent variable Z.

The likelihood for one sample Y^i:



$$P(Y^i, Z|W) = \sum_k \alpha^k P_k(Y^i|W^k)$$

with $\sum_k \alpha^k = 1$. Using Bayes' rule, we can compute the posterior prob of the mixture components for each data point Y^i:

$$r_k(Y^i) = P(Z = k|Y^i, W) = \frac{\alpha^k P_k(Y^i|W^k)}{\sum_j \alpha^j P_j(Y^i|W^j)}$$

These quantities are called "responsabilities".

Learning a Mixture Model with Gradient

We can learn a mixture with gradient descent, but there are much better methods as we will see later. The negative log-likelihood of the data is:

$$L = -\log \prod_i P(Y^i|W) = \sum_i -logP(Y^i|W)$$

Let us consider the likelihood of one data point Y^i:

$$L^i = -logP(Y^i|W) = -log \sum_k \alpha_k P_k(Y^i|W)$$

$$\frac{\partial L^i}{\partial W} = \frac{1}{P(Y^i|W)} \sum_k \alpha_k \frac{\partial P_k(Y^i|W)}{\partial W}$$

Learning a Mixture Model with Gradient (cont)

$$\frac{\partial L^i}{\partial W} = \frac{1}{P(Y^i|W)} \sum_k \alpha_k \frac{\partial P_k(Y^i|W)}{\partial W}$$

$$= \sum_k \alpha_k \frac{1}{P(Y^i|W)} P_k(Y^i|W) \frac{\partial \log P_k(Y^i|W)}{\partial W}$$

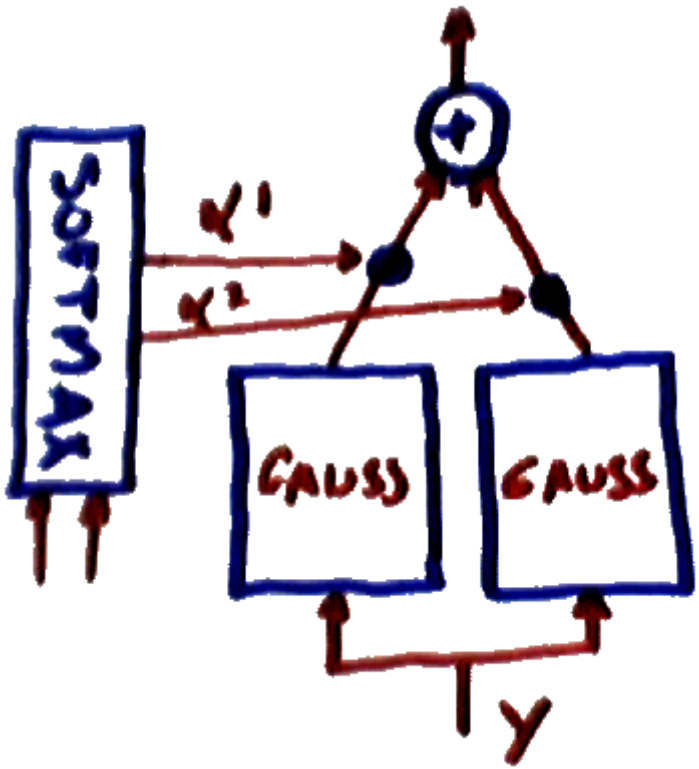
$$= \sum_k \alpha_k \frac{P_k(Y^i|W)}{P(Y^i|W)} \frac{\partial \log P_k(Y^i|W)}{\partial W} == \sum_k r_k(Y^i) \alpha_k \frac{\partial \log P_k(Y^i|W)}{\partial W}$$

The gradient is the weighted sum of gradients of the individual components weighted by the responsabilities.

Example: Gaussian Mixture

$$P(Y|W) = \sum_k \alpha_k |2\pi V^k|^{-1/2} \exp(-1/2(Y - M^k)'(V^k)^{-1}(Y - M^k))$$

This is used a lot in speech recognition.



The Expectation-Maximization Algorithm

Optimizing likelihoods with gradient is the only option in some cases, but there is a considerably more efficient procedure known as EM.
Every time we update the parameters W, the distribution over latent variables Z must be updated as well (because it depends on W.
The basic idea of EM is to keep the distribution over Z constant while we find the optimal W, then we recompute the new distribution over Z that result from the new W, and we iterate. This process is sometimes called *coordinate descent*.

EM: The Trick

The negative log likelihood for a sample Y^i is:

$$L^i = -\log P(Y^i|W) = -\log \int P(Y^i, Z|W) dZ$$

For any distribution $q(Z)$ we can write:

$$L^i = -\log \int q(Z) \frac{P(Y^i, Z|W)}{q(Z)} dZ$$

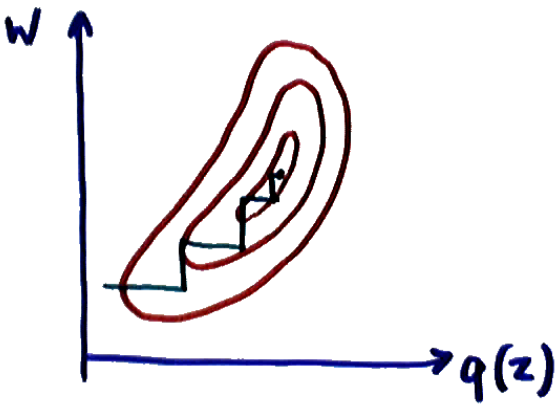
We now use Jensen's inequality, which says that for any concave function G (such as $\log$)

$$-G(\int p(z) f(z) dz) \leq -\int p(z) G(f(z)) dz$$

We get:

$$L^i \leq F^i = -\int q(Z) \log \frac{P(Y^i, Z|W)}{q(Z)} dZ$$

EM



$$L^i \leq F^i = -\int q(Z) \log \frac{P(Y^i, Z|W)}{q(Z)} dZ$$

EM minimizes F^i by alternately
finding the $q(Z)$ that mininizes F **(E-step)**
then finding the W that minimizes F **M-step)**
E-step: $q(Z)^{t+1} \leftarrow \text{argmin}_q F^i(q(Z)^t, W^t)$
M-step: $W(Z)^{t+1} \leftarrow \text{argmin}_W F^i(q(Z)^{t+1}, W^t)$

M Step

We can decompose the free energy:

$$F^i(q(Z), W) = -\int q(Z) \log \frac{P(Y^i, Z|W)}{q(Z)} dZ$$

$$= -\int q(Z) \log P(Y^i, Z|W) dZ + \int q(Z) \log q(Z) dZ$$

The first term is the expected energy with distribution $q(Z)$, the second is the entropy of $q(Z)$, and does not depend on W.
So in the M-step, we only need to consider the first term when minimizing with respect to $q(Z)$.

$$W(Z)^{t+1} \leftarrow \text{argmin}_W - \int q(Z) \log P(Y^i, Z|W) dZ$$

E Step

Proposition: the value of $q(Z)$ that minimizes the free energy is $q(Z) = P(Z|Y^i, W)$
This is the posterior distrib over the latent variabled given teh sample and the current parameter.
Proof:

$$F^i(P(Z|Y^i, W), W) = -\int P(Z|Y^i, W) \log \frac{P(Y^i, Z|W)}{P(Z|Y^i, W)} dZ$$

$$= -\int P(Z|Y^i, W) \log P(Y^i|W) dZ =$$

$$-\log P(Y^i|W) \int_z P(Z|Y^i, W) = -\log P(Y^i|W).1$$