(1) Lecture 4

Note Title 10/7/2006

Generative models $P(I|W)$, $P(W)$.

How to learn $P(W)$?

For simplicity, we will discuss learning a distribution $P(W)$. Replace W by X in this lecture

Ideal Method:

Assume a parameterized model for the distribution of form $p(X|\lambda)$ — λ model parameters.

E.G. Gaussian distribution

$$P(X|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(X-\mu)^2}{2\sigma^2}} \qquad \lambda = (\mu,\sigma).$$

Assume ~~that~~ data is idependent identically distributed (i.i.d).

$$P(X_1,\ldots,X_N|\lambda) = \prod_{i=1}^{N} P(X_i|\lambda).$$

(product for independence).

Choose: $\hat{\lambda} = \arg\max_{\lambda} P(X_1 \ldots X_N|\lambda) = \arg\max \log P(X_1 \ldots X_N|\lambda)$

Hence $p(x_1,\ldots x_N|\hat{\lambda}) \geq p(x_1 \ldots x_N|\lambda)$, for all λ

(2) Example: Gaussian

$$\log P(x_1 \ldots x_N|\mu,\sigma) = \sum_{i=1}^{N} \log P(x_i|\mu,\sigma)$$

$$= -\sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2} - \sum_{i=1}^{N} \log \sqrt{2\pi}\,\sigma.$$

Differentiate w.r.t. μ, σ gives

$$\frac{\partial}{\partial\mu} \log P(x_1 \ldots x_N|\mu,\sigma) = \frac{1}{\sigma^2} \sum_{i=1}^{N} (x_i-\mu).$$

$$\frac{\partial}{\partial\sigma} \log P(x_1 \ldots x_N|\mu,\sigma) = \frac{1}{\sigma^3} \sum_{i=1}^{N} (x_i-\mu)^2 - \frac{N}{\sigma}.$$

Maxima occurs at

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i-\hat{\mu})^2.$$

// Easy to check these are maxima by computing $\frac{\partial^2}{\partial\mu^2}, \frac{\partial^2}{\partial\mu\partial\sigma}, \frac{\partial^2}{\partial\sigma^2}$.

Note: similar results hold for Gaussian distributions in many variables.

Note: The Gaussian is a special case. It is often impossible to solve $\frac{\partial}{\partial\lambda} \log P(x_1 \ldots x_N|\lambda) = 0$ analytically. An algorithm is required.

(3)

Exponential Distributions

$$P(\underline{x}|\underline{\lambda}) = \frac{1}{Z[\underline{\lambda}]} e^{\underline{\lambda} \cdot \underline{\phi}(\underline{x})}$$

normalization factor.

$\underline{\lambda}$ – parameters $\qquad \underline{\lambda} = (\lambda_1, \lambda_2, \dots, \lambda_M)$

$\underline{\phi}(\underline{x})$ – statistics. $\qquad \underline{\phi}(\underline{x}) = (\phi_1(\underline{x}), \phi_2(\underline{x}), \dots, \phi_M(\underline{x}))$.

Almost every named distribution can be expressed as an exponential distribution.

For Gaussian in 1-dimension.

write $\quad \underline{\phi}(x) = (x, x^2) \qquad \underline{\lambda} = \lambda_1, \lambda_2$

$$P(x|\lambda) = \frac{1}{Z[\lambda]} e^{\lambda_1 x + \lambda_2 x^2} \quad \text{compare to} \quad \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Translation

$$\begin{cases} \lambda_2 = -1/2\sigma^2 \\ \lambda_1 = \mu/\sigma^2. \\ Z[\underline{\lambda}] = \sqrt{2\pi}\,\sigma\, e^{\mu^2/2\sigma^2} \end{cases}$$

Similar translation into exponential distributions can be made for Poisson, Beta, Dirichlet ~ most (all) distributions you have been taught.

(4)

Learning an Exponential Distribution

You can learn them by Maximum Likelihood, which again can be interpreted in terms of minimizing the KL divergence between the empirical distribution of the data, and the model distribution.

Examples: $(\underline{x}_1, \underline{x}_2 \dots \underline{x}_\mu \dots \underline{x}_N)$

$$P(\{\underline{x}_1, \underline{x}_2, \dots, \underline{x}_N\} | \underline{\lambda}) = \prod_{\mu=1}^{N} \frac{e^{\underline{\lambda} \cdot \underline{\phi}(\underline{x}_\mu)}}{Z[\underline{\lambda}]}.$$

Maximize wrt $\underline{\lambda}$ ||

This has a very nice form, which occurs because the exponential distribution depends on the data $\underline{x}$ only in terms of the function $\underline{\phi}(\underline{x})$ – the sufficient statistics.

Note:

$$Z[\underline{\lambda}] = \sum_{\underline{x}} e^{\underline{\lambda} \cdot \underline{\phi}(\underline{x})} \qquad \frac{\partial}{\partial \underline{\lambda}} \log Z[\underline{\lambda}] = \sum_{\underline{x}} \frac{\underline{\phi}(\underline{x})\, e^{\underline{\lambda} \cdot \underline{\phi}(\underline{x})}}{Z[\underline{\lambda}]}$$

$$\frac{\partial}{\partial \underline{\lambda}} \log Z[\underline{\lambda}] = \sum_{\underline{x}} \underline{\phi}(\underline{x})\, P(\underline{x} | \underline{\lambda})$$

(5) ML maximizes:

$$\sum_{\mu=1}^{N} \underline{\lambda} \cdot \underline{\phi}(\underline{x}_\mu) - N \log Z[\underline{\lambda}]$$

$$\frac{\partial}{\partial \underline{\lambda}} \rightarrow \quad \sum_{\mu=1}^{N} \underline{\phi}(\underline{x}_\mu) - N \sum_{\underline{x}} \underline{\phi}(\underline{x}) P(\underline{x}|\underline{\lambda}).$$

$$\sum_{\underline{x}} \underline{\phi}(\underline{x}) P(\underline{x}|\underline{\lambda}) = \frac{1}{N} \sum_{\mu=1}^{N} \underline{\phi}(\underline{x}_\mu)$$

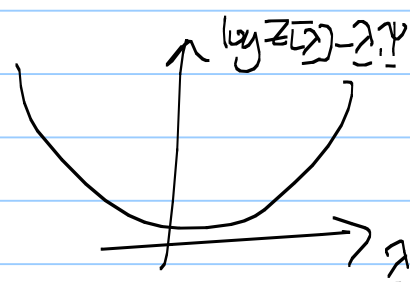
Pick the parameters $\underline{\lambda}$ so that the average of the statistics $\underline{\phi}(\underline{x})$ w.r.t distribution $P(\underline{x}|\underline{\lambda})$ is equal to the average of the statistics of the samples.

Solve: $\sum_{\underline{x}} \underline{\phi}(\underline{x}) P(\underline{x}|\underline{\lambda}) = \underline{\psi}$

with $\underline{\psi} = \frac{1}{N} \sum_{\mu=1}^{N} \underline{\phi}(\underline{x}_\mu)$,

This is equivalent to minimizing

$$\log Z[\underline{\lambda}] - \underline{\lambda} \cdot \underline{\psi}$$



It can be shown that this function is convex and has a unique solution:

(Because $\frac{\partial^2}{\partial \underline{\lambda} \partial \underline{\lambda}} \{\log Z[\underline{\lambda}] - \underline{\lambda} \cdot \underline{\psi}\}$ is positive definite).

(6) ML estimation for exponential distributions is a convex optimization function — this means ~~that~~ their are algorithms which are guaranteed to converge to the correct solution

<u>Example</u>: Generalized Iterative Scaling (GIS)

$$\begin{cases} \underline{\lambda}^{t+1} = \underline{\lambda}^{t} - \log \underline{\Psi}^{t} + \log \underline{\Psi} \\ \text{where} \quad \underline{\Psi}^{t} = \sum_{\underline{x}} \underline{\phi}(\underline{x})\, P(\underline{x} \mid \underline{\lambda}^{t}). \\ \text{Notation:} \quad \log \underline{\Psi} \text{ is a vector with components } \log \Psi_1, \log \Psi_2, \ldots, \log \Psi_M \end{cases}$$

<u>But</u> this requires computing

$$\sum_{\underline{x}} \underline{\phi}(\underline{x})\, P(\underline{x} \mid \underline{\lambda}^{t})$$

which is often difficult.

(For people who took CS 202, there are often statistical/MCMC methods which can (approximately) compute this rapidly.

(7) How does this apply to vision?

w_i w_j

Consider the weak membrane model of images

Geman & Geman, Blake & Zisserman, Mumford & Shah.

In probabilistic terms, this can be formulated as

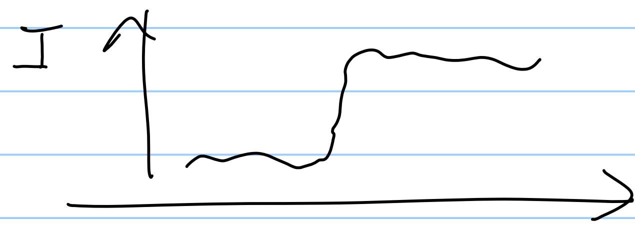
$$P(I|W) = \prod_i P(I_i|W_i)$$

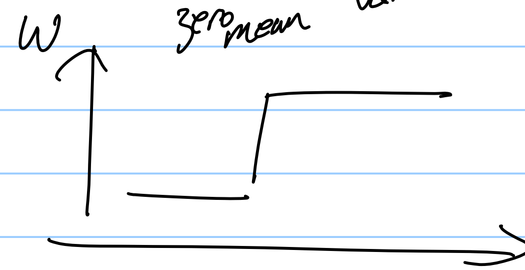
with $P(I_i|W_i) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(I_i - W_i)^2/2\sigma^2}$

The observed image I is a corrupted version of a true image W. The corruption is by additive Gaussian noise

$$\underbrace{I_i = W_i + \epsilon_i}_{\text{noise}} \qquad \epsilon_i \sim N(0, \sigma)$$

zero mean, variance





Need a prior $P(W)$ for the 'ideal image'.

Gibbs Distribution $\quad P(W) = \frac{1}{Z} e^{-E[W]}$

$$E[W] = \sum_i \sum_{j \in Nbd(i)} \psi(W_i, W_j)$$

(8) $Nbd(i)$ is the neighborhood structure

e.g. image lattice.

What function $\psi(w_i, w_j)$?

A natural choice is $\psi(w_i, w_j) = (w_i - w_j)^2$

penalize the square of the difference between the intensities of neighboring pixels.

Advantages:

(i) this makes it easy to learn the distribution from training data. It is a Gaussian distribution which, as we have seen, can be learnt by analytic methods (i.e. no need for steepest descent or GIS.)

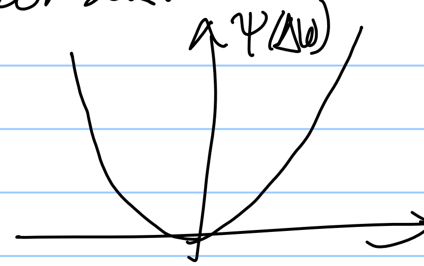
(ii) this makes inference easy. To estimate $\hat{w} = \text{ARG MAX } P(W|I)$

reduces to minimizing.

$$E[W] = \frac{1}{2\sigma^2}\sum_i (w_i - I_i)^2 + \sum_i \sum_{j \in N(i)} \psi(w_i, w_j)$$

If $\psi(w_i, w_j)$ is quadratic — $(w_i - w_j)^2$ — then $E[W]$ is quadratic, and its minimum can be found by solving linear equations $\frac{\partial E}{\partial w} = 0$

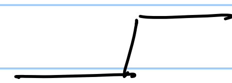
(9) But although $(w_i - w_j)^2$ is good for inference and learning it is not a good prior distribution.

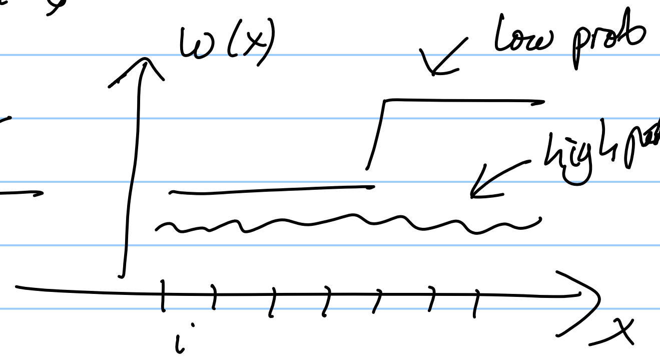


It penalizes large $\Delta w = |w_i - w_j|$ too much.

It penalizes small $\Delta w = |w_i - w_j|$ too little.

It prefers images like 

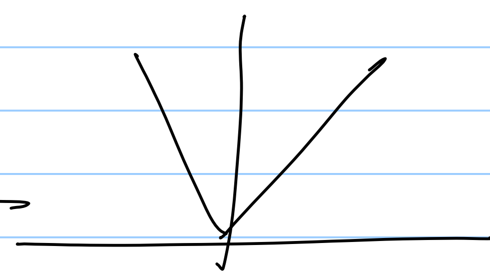
and distelikes images like 

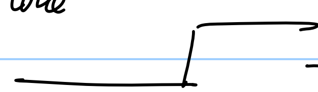


A better prior is $\varphi(w_i, w_j) = |w_i - w_j|$

L^1 norm.

this prefers images that are very smooth 



it discourages images like 

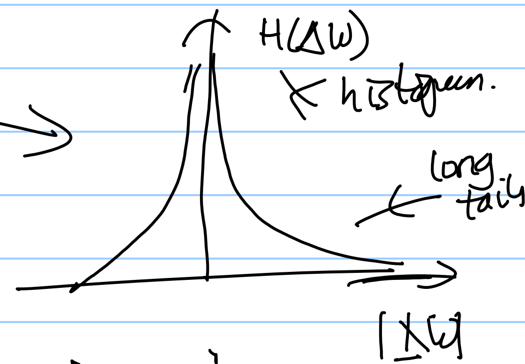
but tolerates them better that the quadratic penalty.

<u>Now</u>, try to learn the prior from data.

(10)

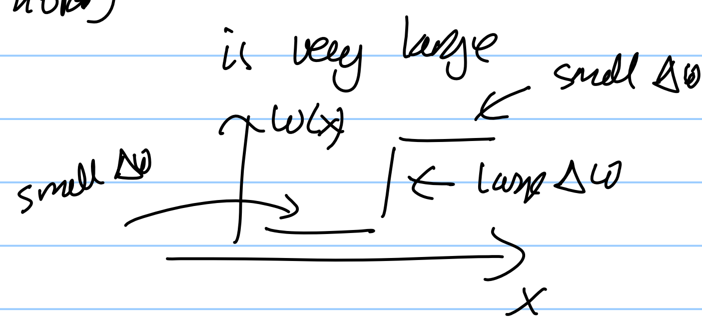
Observation, plot the histogram of $\Delta W = (W_i - W_j) \quad j \in \mathcal{N}(i)$ for image pixels.

This takes a standard form



This is inconsistent with a Gaussian distribution. → the distribution is peaked at 0 (i.e. the intensity of most pixels is similar to those of their neighbors)

but ΔW can also be large – i.e. at edges.



Measure the histogram

$$H(\alpha) = \sum_i \sum_{j \in \mathcal{N}(i)} \delta(\alpha, W_i - W_j)$$

$\delta(z, W_i - W_j) = 1$, if $z = W_i - W_j$
$= 0$, otherwise.

Exponential distribution (max-entropy principle) next lecture.

$$P(W) = \frac{1}{Z} e^{-\sum_\alpha \lambda(\alpha) H(\alpha)}$$

$$\sum_\alpha \lambda(\alpha) H(\alpha) = \sum_\alpha \lambda(\alpha) \sum_i \sum_{j \in \mathcal{N}(i)} \delta(\alpha, W_i - W_j)$$

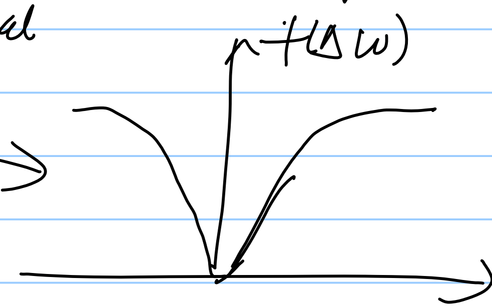
$$= \sum_i \sum_{j \in \mathcal{N}(i)} \lambda(W_i - W_j).$$

(11) Here $P(W) = \frac{1}{Z} e^{-\sum_i \sum_{j \in Nbh(i)} \lambda(w_i - w_j)}$

Thus, we obtain a probability distribution with the same form of the potentials — i.e. $\sum_i \sum_{j \in Nbh(i)} \phi(w_i, w_j)$ — by assuming an exponential form for the distribution and choosing the statistic $H(\alpha) = \sum_i \sum_{j \in Nb(i)} \delta(\alpha, w_i - w_j)$.

(Zhu & Mumford 1997)

The form of the learnt potential is



This was suggested earlier (Geman & Geman) by specifying additional "line process" variables.

$$P(W) \rightarrow P(W, L) = \frac{1}{Z} e^{-E(W,L)}$$

$$E(W,L) = \sum_i \sum_{j \in N(i)} (w_i - w_j)^2 (1 - L_{ij}) + K \sum_i \sum_{j \in N(i)} L_{ij}$$

$L_{ij} \in \{0, 1\}$ is a binary-valued variable. $L_{ij} = 1$ 'cuts' the smoothness between w_i & w_j.

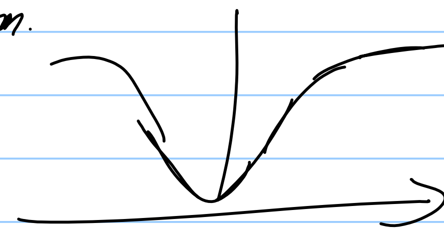
Lower Energy (higher probability) if $L_{ij} = 1$, when $|w_i - w_j| > K^{1/2}$
if $L_{ij} = 0$, when $|w_i - w_j| < K^{1/2}$

(12)

$$P(w,L) = \frac{1}{Z} e^{-E(w,L)}$$

$$P(w) = \sum_L P(w,L) = \frac{1}{Z} e^{-\sum_i \sum_{j \in Nbh(i)} \phi(w_i, w_j)}$$

It can be shown that $\varphi(w_i, w_j)$ has form – similar to the least form.



Note: more accurate priors can be obtained by considering higher order filters $\varphi(w_i, w_j, \ldots w_k)$

It can be shown that the statistics of any filter. $\sum_i A_i w_i$, such that $\sum_i A_i = 0$ has very similar form on each image (M. Green, Maths, UCLA)

But this leads to complicated probability distributions.

Hard to do inference on them.