LECTURE NOTE #4

PROF. ALAN YUILLE

1. Learning Probability Distributions (Parametric methods)

$p(x \mid y)$ & $p(y)$

For simplicity, we will discuss learning a distribution $p(x)$.

Ideal Method

Assume a parameterized model for the distribution of form $p(x \mid \theta)$, $\quad \theta :$ model parameter

E.G.

Gaussian distribution

$p(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, $\quad \theta = (\mu, \sigma)$

Assume

that data is independent identically distributed (iid).

$p(x_1, \ldots, x_N \mid \theta) = \prod_{i=1}^{N} p(x_i \mid \theta)$ $\quad$ (product for independence).

Choose:

$\hat{\theta} = \arg_\theta \max p(x_1, \ldots, x_N \mid \theta) = \arg_\theta \min\{-\log p(x_1, \ldots, x_N \mid \theta)\}$ (use $\log\{a \times b\} = \log a + \log b$).

Hence $p(x_1, \ldots, x_N \mid \hat{\theta}) \geq p(x_1, \ldots, x_N \mid \theta)$, for all θ

2.

Example: Gaussian

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

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

Differentiate w,r,l. μ, σ gives

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

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

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 the second order derivatives (Hessian) and showing it is positive definite. Hence the (negative) log likelihood is a convex function and has at most one minimum.

$\frac{\delta^2}{\delta\mu^2}, \frac{\delta^2}{\delta\mu\delta\sigma}, \frac{\delta^2}{\delta\sigma^2}$

Note:

Similar results hold for Gaussian distribution in higher dimensions.

Note:

The Gaussian is a special case. It is often impossible to some $\frac{\delta}{\delta\theta} \log p(x_i, \ldots, x_N \mid \theta) = 0$ analytically. An algorithm is required (see later).

3.

An alternative viewpoint on ML learning of distributions. This gives deeper understanding.
Suppose the data is generated by a distribution $f(x)$.

Define the Kullback-Leiber divergence between $f(x)$ and the model $p(x|\theta)$

Kullback-Leiber: $D(f||p) = \sum_x f(x) \log \frac{f(x)}{p(x|\theta)}$

KL has the property that
$D(f||p) \geqq 0 \quad \forall f, p$
$D(f||p) = 0,$ if, and only if, $f(x) = p(x|\theta)$

So, $D(f||p)$ is a measure of the similarity between $f(x)$ and $p(x|\theta)$

We can write, $D(f||p) = \sum_x f(x) \log f(x) - \sum_x f(x) \log p(x|\theta)$
* $\sum_x f(x) \log f(x)$: Independent of θ
* $\sum_x f(x) \log p(x|\theta)$: Depends on θ

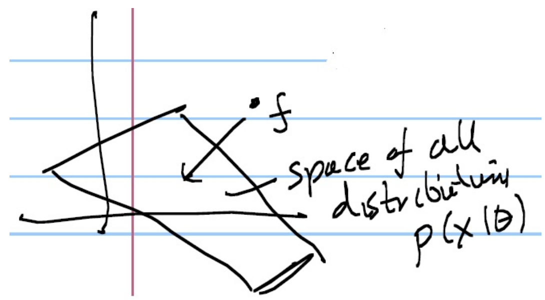

FIGURE 1. space of all distribution $p(x|\theta)$ in section 3.

4.

Now suppose we have sample (i.i.d.) $x_1, ..., x_n$ from $f(x)$

This gives us on empirical distribution
$f_{emp}(x) = \frac{1}{N}\sum_{i=1}^{N}\delta_{x,x_i}$

* δ_{x,x_i}: Kronecker delta – an Indicator function

The KL divergence between $f_{emp}(x)$ and $p(x|\theta)$ can be written as:

$J(\theta) = -\sum_x f_{emp}(x)\log p(x|\theta) + K$ * K is independent of θ
$J(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\log p(x_i|\theta) + K$

Minimizing $J(\theta)$ w.r.t. θ, finds the distribution $p(x|\hat{\theta})$ which is closest to $f_{emp}(x)$.

But minimizing $J(\theta)$ w.r.t. θ is exactly ML.
$\hat{\theta} = \arg\min_\theta\{-\sum_{i=1}^{N}\log p(x_i|\theta)\}$

So, ML has meaning even if best fit to the model. Even if the model is only an approximation.

5. Exponential Distributions

$p(\underline{x}|\underline{\lambda}) = \frac{1}{Z[\underline{\lambda}]} \exp^{\underline{\lambda}\cdot\underline{\phi}(\underline{x})}$
* $z[\underline{\lambda}]$: normalization factor
* $\underline{\lambda}$: parameters $\lambda = (\lambda_1, \lambda_2, ..., \lambda_M)$
* $\underline{\phi}(\underline{x})$: statistics $\underline{\phi}(\underline{x}) = (\phi_1(\underline{x}), \phi_2(\underline{x}), ..., \phi_M(\underline{x}))$

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

For Gaussian in 1-dimension
write $\underline{\phi}(x) = (x, x^2)$ $\underline{\lambda} = (\lambda_1 \lambda_2)$
$p(x|\lambda) = \frac{1}{z[\underline{\lambda}]} \exp^{\lambda_1 x + \lambda_2 x^2}$ compare to $\frac{1}{\sqrt{2\pi}\sigma} \exp^{\frac{-(x-\mu)^2}{2\sigma^2}}$

Translation

$$\begin{cases} \lambda_2 = -\frac{1}{2\sigma^2} \\ \lambda_1 = \frac{\mu}{\sigma^2} \\ Z[\underline{\lambda}] = \sqrt{2\pi}\sigma \exp \frac{\mu^2}{2\sigma^2} \end{cases}$$

Similar translations into exponential distribution can be made for Poisson, Beta, Dirichlet $\sim$ most (all) distribution you have been taught.

6. 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.

Example:

$(\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_\mu, \ldots, \underline{x}_N,)$

$p(\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_N \mid \underline{\lambda}) = \quad \prod_{\mu=1}^{N} e^{\frac{\underline{\lambda} \cdot \underline{\phi}(\underline{x}_\mu)}{Z[\underline{\lambda}]}}$

Maximize w.r.t $\underline{\lambda}$:||

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

Note :

$Z[\underline{\lambda}] = \sum_{\underline{x}} e^{\underline{\lambda} \cdot \underline{\phi}(\underline{x})}$

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

$\frac{\delta}{\delta \underline{\lambda}} \log Z[\underline{\lambda}] = \sum_{\underline{x}} \underline{\phi}(\underline{x}) p(\underline{x} \mid \underline{\lambda})$

7.

ML

minimizes :

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

$\frac{\delta}{\delta \underline{\lambda}} \longrightarrow -\sum_{\mu=1}^{N} \underline{\lambda} \cdot \underline{\phi}(\underline{x}_{\mu}) + N \sum_{\underline{x}} \underline{\phi}(\underline{x}) p(\underline{x} \mid \underline{\lambda})$

$\sum_{\underline{x}} \underline{\phi}(\underline{x}) p(\underline{x} \mid \underline{\lambda}) = \frac{1}{N} \sum_{\mu=1}^{N} \underline{\phi}(\underline{x_{\mu}})$

Pick the parameters $\underline{\lambda}$ so that the *expected statistics* $\underline{\phi}(\underline{x})$ *with respect to the distribution* $p(\underline{x} \mid \underline{\lambda})$ *is equal to the average of the statistics of the samples.*

This requires us to solve:

$\sum_{\underline{x}} \underline{\phi}(\underline{x}) p(\underline{x} \mid \underline{\lambda}) = \underline{\psi}$ with $\underline{\psi} = \frac{1}{N} \sum_{\mu}^{N} \underline{\phi}(\underline{x}_{\mu})$.

This is equivalent to minimizing.

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

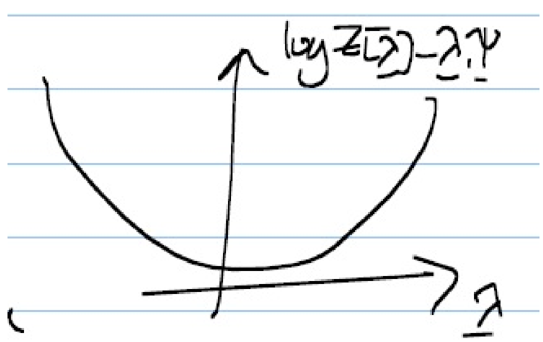

FIGURE 2. $\log z[\underline{\lambda}] - \underline{\lambda} \cdot \underline{\psi}$ in section 7.

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

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

8.

ML estimation for exponential distributions is a convex optimization function - this means that there are algorithms which are guaranteed to converge to the correct solution.

Example:

Generalized Iterative Scaling (GIS)
Initialize $\lambda^{t=0}$ to any value. Then iterate:

$$\begin{cases} \underline{\lambda}^{t+1} = \underline{\lambda}^{t} - \log \underline{\psi}^{t} + \log \underline{\psi} \\ \text{where} \quad \underline{\psi}^{t} = \sum_{\underline{x}} \underline{\phi}(\underline{x}) p(\underline{x} \mid \underline{\lambda}^{t}) \\ \text{Notation : } \log \underline{\psi} \text{ is a vector with components} \quad \log \psi_1, \log \psi_1, \ldots, \log \psi_N \end{cases}$$

This algorithm is guaranteed to converge to the correct solution for any starting point $\lambda^{t=0}$ (because $\log Z[\vec{\lambda}] - \vec{\lambda} \cdot \vec{\psi}$ is convex). If it reaches a value $\vec{\lambda}$ such that $\log \underline{\psi}^{t} = \log \underline{\psi}$ – i.e. the expected statistics of the model equals the statistics of the data – then the algorithm stops – $\vec{\lambda}^{t+1} = \vec{\lambda}^{t}$.

But

the algorithm requires computing the quantity

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

for each iteration step, which is often difficult (see examples in the next lecture).

Note: Markov Chain Monte Carlo (MCMC) algorithms can be used to approximate this term.

9. The Maximum Entropy Principle

LECTURE NOTE #4

How to get to distributions from statistics.

Suppose we measure some statistics $\underline{\phi}(\underline{x})$, what distribution does it correspond to? Impossible question. There are too many possible distributions.

Maximum Entropy Principle:
Select the distribution which has the maximum entropy and is consistent with the observed statistics

Entropy of a distribution $p(\underline{x})$
$H[p] = -\sum_x p(\underline{x}) \log p(\underline{x})$

A measure of the amount of information obtain by observing a sample $\underline{x}$ from a distribution $p(\underline{x})$.

Shannon - Information Theory. Encode a signal x by a code of length $-\log p(x)$ – so that frequent signals ($p(x)$ big) have short codes and infrequent signals ($p(x)$ small) have long codes. Then the expected code length is $-\sum_x p(x) \log p(x)$. Alternatively, the entropy is the amount of information we expect to get from a signal x before we observe it – but we know that the signal has been sampled from a distribution $p(x)$.
Entropy is a concept discovered by physicists. It can be shown that the entropy of a physical system always increases (with plausible assumptions). This is called the Second Law of Thermodynamics. It explains why a cup can break into many pieces (if you drop it), but a cup can never be created by its pieces suddenly joining together. Thermodynamics was discovered in the early 19^{th} century, and shows that it is impossible to design an engine that can create energy.

10.

Example: Suppose $\underline{x}$ can take N values: $\underline{\alpha}_1, \underline{\alpha}_2, ..., \underline{\alpha}_N$

Suppose:
$p(\underline{x} = \underline{\alpha}_1) = 1$

$p(\underline{x} = \underline{\alpha}_j) = 0, \qquad j = 2, ..., N$

Then the entropy of this distribution is zero, because we know that $\underline{x}$ has to take value α, before we observe it. The entropy is $-0 \log 0 + (N-1)\{1 log 1\}$, and $0 \log 0 = 0$ and $1 \log 1 = 0$ (take the limit of $x \log x$ as $x \mapsto 0$ and $x \mapsto 1$.)

Now suppose:
$p(\underline{x} = \underline{\alpha}_j) = \frac{1}{N}, \qquad j = 1, ..., N$
Then $H(p) = -N \times \frac{1}{N} \log(\frac{1}{N}) = \log N$

This is the maximum entropy distribution. Note that the maximum entropy distribution is *uniform* – all states x are equally likely.

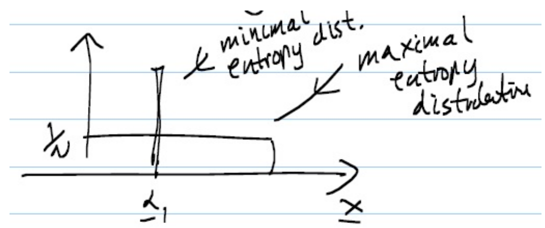

FIGURE 3. maximum entropy distribution in section 10.

11. Maximum Entropy Principle

Given statistics $\phi(\underline{x})$ with observed value $\underline{\psi}$, choose the distribution $p(\underline{x})$ to maximize the entropy subject to constraints (Jaynes).

$-\sum_{\underline{x}} p(\underline{x}) \log p(\underline{x}) + \mu\{\sum_{\underline{x}} p(\underline{x}) - 1\} + \underline{\lambda} \cdot \{\sum_{\underline{x}} p(\underline{x})\phi(\underline{x}) - \underline{\psi}\}$

μ, λ: lagrange multipliers $\quad p(\underline{x})$: constraints

$\frac{\delta}{\delta p(\underline{x})} \qquad -\log p(\underline{x}) - 1 + \mu + \underline{\lambda} \cdot \underline{\phi}(\underline{x}) = 0$

Solution, $p(\underline{x}|\underline{\lambda}) = \frac{\exp^{\underline{\lambda}\cdot\underline{\phi}(\underline{x})}}{Z[\underline{\lambda}]}$

where $\underline{\lambda}, Z[\underline{\lambda}]$ are chosen to satisfy the constraints:

$\sum_{\underline{x}} p(\underline{x}) = 1, \Rightarrow Z[\underline{\lambda}] = \sum_{\underline{x}} \exp^{\underline{\lambda}\cdot\underline{\phi}(\underline{x})}$
$\sum_{\underline{x}} p(\underline{x})\phi(\underline{x}) = \underline{\psi}, \Rightarrow \underline{\lambda}$ is chosen s.t. $\sum_{\underline{x}} p(\underline{x}|\underline{\lambda})\phi(\underline{x}) = \underline{\psi}$

The maximum entropy principle recovers exponential distribution!

12. Training and Testing

Critical issue is - how much data do you need to learn a distribution?

There is no perfect answer. A rule of thumb is that you need $k\times$ no. of parameter of the distribution, where $k = 5to10$.

In practice, train (learn) the model on a training dataset. Test it on a second dataset. $\forall$ performance (e.g. Bayes Risk, ROC curves, etc) is the some on both - then you have learned or generalized

$\forall$ performance is good on the training set but bad on the training set - then you have only memorized the training set.