(1)
Alpaydin
Chp 4.

Parametric Methods

3/26/2008

How to estimate the probabilities to make decisions (previous lecture).

Abstract: estimate the sufficient statistics which are functions of the data samples. These are sufficient to determine the parameters of the distribution.

Density estimation $p(\underline{x})$
Use this to estimate $p(\underline{x}|C_i)$ & $p(C_i)$

Maximum Likelihood Estimation (MLE)

Sample $X = \{x^t\}_{t=1}^N$
i.i.d. from some distribution $p(x|\theta)$ ← parameter.

likelihood $\ell(\theta) = p(X|\theta) = \prod_{t=1}^N p(x^t|\theta)$

MLE: Find $\hat{\theta}$ to maximize $\ell(\theta)$
or, equivalently maximize the log likelihood
$$L(\theta|\underline{X}) = \log \ell(\theta|X) = \sum_{t=1}^N \log p(\underline{x}^t|\theta)$$

(2)

Bernoulli Density

two outcomes,
random variable X $\quad$ $X=1$, with prob p
$X=0$, with prob $1-p$.

$$p(x) = p^x (1-p)^{1-x}, \quad x \in [0,1].$$

p is the only parameter.

Log-likelihood

$$\mathcal{L}(p|\mathcal{X}) = \log \prod_{t=1}^{N} p^{x^t} (1-p)^{1-x^t}$$

$$= \sum_t x^t \log p + \left(N - \sum_t x^t\right) \log(1-p).$$

Solve $\frac{\partial L(p|\mathcal{X})}{\partial p} = 0$ gives $\hat{p} = \frac{\sum_t x^t}{N}$.

Intuitive, no. of positive outcomes
divided by number of samples.

Multinomial Density

generalization of Bernoulli
K classes, probability p_i, $\sum_{i=1}^{K} p_i = 1$

$$p(x_1, x_2 \ldots x_K) = \prod_{i=1}^{K} p_i^{x_i}$$

x_i's indicator variables $\quad$ $x_i = 1$, if outcome is i
0, otherwise.

(3) N experiments with outcomes $\mathcal{X} = \{\underline{x}^t\}_{t=1}^N$

$x^t_i = \begin{cases} 1, & \text{if experiment } t \text{ choose state } t \\ 0, & \text{otherwise} \end{cases}$

MLE of p_i is $\hat{p}_i = \sum_t x^t_i / N$

proportion of sample which lie in class i.

Gaussian (Normal) Density

$$p(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad -\infty < x < \infty$$

$$\mathcal{N}(\mu, \sigma^2)$$

Sample $\quad \mathcal{X} = \{x^t\}_{t=1}^N$, with $x^t \sim \mathcal{N}(\mu, \sigma^2)$

log likelihood $\quad \mathcal{L}(\mu, \sigma \mid \mathcal{X}) = -\frac{N}{2} \log 2\pi - N \log \sigma - \sum_t \frac{(x^t - \mu)^2}{2\sigma^2}$.

MLE
$$m = \sum_t x^t / N$$
$$s^2 = \sum_t \frac{(x^t - m)^2}{N}$$

Convention: Greek for the parameters.
Roman for their estimates.

Exponential Distributions and Maximum Likelihood

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

exponential distribution
statistics $\phi(\cdot)$
parameters λ

$$Z[\lambda] = \sum_{x} e^{\lambda \cdot \phi(x)}, \text{ if } x \text{ is discrete-valued}$$

normalization term

$$= \int dx\, e^{\lambda \cdot \phi(x)}, \text{ if } x \text{ is continuous-valued}$$

Ensures $\sum_{x} P(x|\lambda) = 1$, or $\int dx\, P(x|\lambda) = 1$.

Note: (1) most "named" distributions can be expressed as exponential distribution $\Rightarrow$ e.g. Gaussian, Poisson, Bernoulli, ...

(2) any distribution can be approximated arbitarily accurately by an exponential distribution.

(3) mixture models can be expressed as exponential distribution by introducing hidden variables

Example: Gaussian in 1-D $P(x:\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$

$$\begin{cases} \log P(x:\mu,\sigma) = -\frac{1}{2\sigma^2}x^2 + \frac{x\mu}{\sigma^2} - \frac{1}{2\sigma^2}\mu^2 - \log\sqrt{2\pi}\sigma \\ \log P(x:\lambda) = \lambda \cdot \phi(x) - \log Z[\lambda] \end{cases}$$

s.t. $\phi(x) = (x, x^2)$

$\lambda = \left(-\frac{1}{2\sigma^2}, \frac{\mu}{\sigma^2}\right)$

$$\log Z[\lambda] = -\frac{1}{2\sigma^2}\mu^2 - \log\sqrt{2\pi}\sigma$$

[5]

Note: the sufficient statistics $\phi(x)$ are the properties of the data x that we need to store (we can throw out the rest of the data).

For a set of samples $x^1, x^2, \ldots x^M$ from a Gaussian
we only need to store the statistics

$$\frac{1}{M}\sum_{i=1}^{M}(x^i) \quad \text{and} \quad \frac{1}{M}\sum_{i=1}^{M}(x^i)^2.$$

the parameters of the Gaussian — μ & σ^2 — estimated by ML will depend only on these statistics.
(e.g. we don't need to also store $\frac{1}{M}\sum_{i=1}^{M}(x^i)^4$, or even all the samples $x^1, x^2, \ldots, x^M$)

Another example: Bernoulli.

$$p(x) = \theta^x (1-\theta)^{1-x}$$

$$\log p(x) = x\log\theta + (1-x)\log(1-\theta)$$

$$= x\{\log\theta - \log(1-\theta)\} + \log(1-\theta)$$

Exponential

$$\phi(x) = x$$

$$\lambda = \log\theta - \log(1-\theta)$$

$$\log Z[\lambda] = \log(1-\theta)$$

(6) Maximum Likelihood for Exponential Distributions

Samples $\{x^i : i = 1 \text{ to } M\}$ from $P(x|\lambda) = \frac{e^{\underline{\lambda} \cdot \underline{\phi}(x)}}{Z[\lambda]}$

ML says maximize w.r.t. λ $\prod_{i=1}^{M} P(x^i|\lambda)$

$$\hat{\underline{\lambda}} = \arg\max_{\lambda} \sum_{i=1}^{M} \log P(x^i|\lambda)$$

$$\hat{\underline{\lambda}} = \arg\max_{\lambda} \{ \underline{\lambda} \cdot \sum_{i=1}^{M} \underline{\phi}(\underline{x}^i) - M \log Z[\lambda] \}$$

Equivalently: minimize $F[\underline{\lambda}] = \log Z[\lambda] - \underline{\lambda} \cdot \frac{1}{M} \sum_{i=1}^{M} \underline{\phi}(\underline{x}^i)$

$$F[\lambda] = \log Z[\lambda] - \underline{\lambda} \cdot \underline{\Psi}$$

where $\underline{\Psi} = \frac{1}{M} \sum_{i=1}^{M} \underline{\phi}(\underline{x}^i)$ is the empirical statistics

i.e.g. $\left(\frac{1}{M} \sum_{i=1}^{M} x^i, \frac{1}{M} \sum_{i=1}^{M} (x^i)^2 \right)$ for Gaussian)

Minimize: solve $\frac{\partial F[\lambda]}{\partial \underline{\lambda}} = 0$, for $\underline{\lambda}$:

Note $F[\lambda]$ is convex — $\frac{\partial^2 F}{\partial \underline{\lambda} \partial \underline{\lambda}} \geq 0$ by properties of $\log Z[\lambda]$

so there is a unique solution

(property of $\log Z[\lambda]$)
$$\frac{\partial F}{\partial \underline{\lambda}} = \frac{\partial \log Z[\lambda]}{\partial \underline{\lambda}} - \underline{\Psi} = \sum_{x} P(x|\lambda) \underline{\phi}(\underline{x}) = \underline{\Psi}$$

Maximum Likelihood selects $\underline{\lambda}$ for that the expected statistics $\sum_x P(x|\lambda)\underline{\phi}(x)$ equal the observed statistics $\underline{\Psi}$.

— e.g. for a Gaussian the mean and variance parameters λ, μ are chosen to equal the mean and variance of the data

(7)

Important:

how to solve $\sum_x P(x|\lambda)\underline{\phi}(x) = \underline{\Psi}$?

(or $\int d\underline{x}\, P(\underline{x}|\lambda)\underline{\phi}(\underline{x}) = \underline{\Psi}$) ?

For special distributions — like Gaussians, Bernoulli, Poisson — these can be solved analytically to give a closed form solution

→ e.g. $\mu = \frac{1}{N}\sum_{i=1}^{N} x^i$ for Gaussian

$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x^i)^2 + \mu^2$

But for other distributions we cannot solve analytically. (or Generalized Iterative Scaling GIS)

In these cases we can do steepest descent on $F[\underline{\lambda}]$ — which is convex, so steepest descent converges to the unique solution.

$$\underline{\lambda}^{t+1} = \underline{\lambda}^t - \epsilon \frac{\partial F[\underline{\lambda}]}{\partial \underline{\lambda}}\Big|_{\underline{\lambda}^t}$$

$t=0$
t — index of iterations.

$$\underline{\lambda}^{t+1} = \underline{\lambda}^t - \underbrace{\epsilon \sum_{\underline{x}} \underline{\phi}(\underline{x}) P(\underline{x}:\underline{\lambda}^t) + \epsilon\underline{\Psi}}$$

converges — when $\sum_{\underline{x}} \underline{\phi}(\underline{x}) P(\underline{x}:\underline{\lambda}^t) = \underline{\Psi}$

model statistics become the empirical statistics

Problem: each iteration requires estimating $\sum_{\underline{x}} \underline{\phi}(\underline{x}) P(\underline{x}:\underline{\lambda}^t)$, which is difficult — note, if we can compute this analytically then we can probably solve $\sum_{\underline{x}} \underline{\phi}(\underline{x}) P(\underline{x}:\underline{\lambda}) = \underline{\Psi}$ directly (e.g. for Gaussian $\int_{\underline{x}} \underline{\phi}(\underline{x}) P(\underline{x}:\underline{\lambda}^t) d\underline{x} = (\mu, \sigma^2 + \mu^2)$)

Alternative perspective: Maximum Entropy

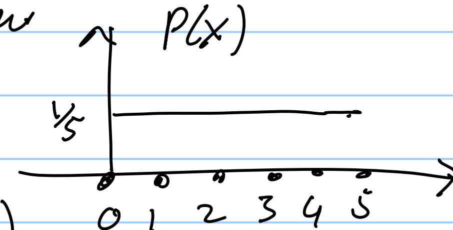
Entropy of a distribution

$$-\sum_x p(x|\lambda) \log p(x|\lambda) \quad , \text{or} -\int dx\, p(x|\lambda) \log p(x|\lambda)$$

expected information from a sample x of a distribution

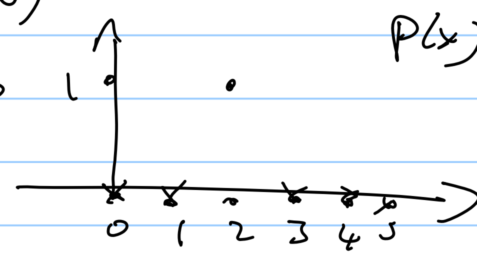
Shannon's Theory of Information

uniform distribution



entropy $-5\left(\frac{1}{5}\right)\log(1/5) = \log 5$

generally $\log N$ (if N no. of states)



entropy $-4(0 \log 0) - (1 \log 1)$
$= 0$

$\left(\begin{array}{l} 0 \log 0 = 0 \\ 1 \log 1 = 0 \end{array}\right)$

no information is transmitted if we get a sample from this distribution, since we know what the value will be before we see it (it has to be 2)

But we do get information from a sample from the uniform distribution

Shannon; says we should encode a sample x from $p(x)$ by $-\log p(x)$ bits — so samples with high probability have low number of bits — then the expected coding cost (in bits) is just the entropy

$$\sum_x p(x)\{-\log p(x)\} = -\sum_x p(x) \log p(x)$$

(page 9) <u>Maximum entropy principle:</u>

suppose that we observe some statistics $\phi(x)$ and they take a value $\underline{\Psi}$

(e.g. the mean of the data takes a value).

<u>What</u> distribution describes the data?

Maximum Entropy Principle (see Jaynes)

Suggest we select the distribution $P(x)$ that maximizes the entropy (i.e. is as uniform as possible) but is consistent with the data

$$L(p:\lambda,\mu) = -\sum_x p(x)\log p(x) + \mu\{\sum_x p(x) - 1\} + \underline{\lambda}\cdot\{\sum_x p(x)\,\underline{\phi}(\underline{x}) - \underline{\Psi}\}$$

Lagrange multipliers

Lagrange multipliers impose the constraint

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

$$\Rightarrow p(\underline{x}) = e^{\underline{\lambda}\cdot\underline{\phi}(\underline{x})}\,e^{-1+\mu}. \quad (*)$$

$$\frac{\partial L}{\partial \mu} = 0 \Rightarrow \sum_x p(x) = 1, \quad \text{normalization}$$

$$\frac{\partial L}{\partial \underline{\lambda}} = 0 \Rightarrow \sum_x p(x)\,\underline{\phi}(\underline{x}) = \underline{\Psi}, \quad \text{consistency with the data}$$

Imposing normalization on (*) gives

$$P(\underline{x}:\underline{\lambda}) = \frac{1}{Z[\underline{\lambda}]}\,e^{\underline{\lambda}\cdot\underline{\phi}(\underline{x})} \rightarrow \underline{\text{exponential distribution}}!$$

$$\text{s.t.} \quad \sum_x P(\underline{x}:\underline{\lambda})\,\underline{\phi}(\underline{x}) = \underline{\Psi}.$$

<u>So</u> Maximum Likelihood with Exponential Distribution is the same as doing Maximum Entropy with the statistics

(10) Maximum Entropy suggests a solution to the problem — how do we know what distribution to use?

Maximum Entropy strategy is to search over the statistics — instead of searching over distributions

i.e. statistic $\underline{\phi}_1(.)$ gives distribution $P(\underline{x}) = \frac{1}{Z[\underline{\lambda}_1]} e^{\underline{\lambda}_1 \cdot \underline{\phi}_1(\underline{x})}$

" $\underline{\phi}_2(\,)$ " " $P(\underline{x}) = \frac{1}{Z[\underline{\lambda}_2]} e^{\underline{\lambda} \cdot \underline{\phi}(\underline{x})}$

" $\underline{\phi}_1(x)$ & $\underline{\phi}_2(\underline{x})$ " " $P(\underline{x}) = \frac{1}{Z[\underline{\lambda}_1, \underline{\lambda}_2]} e^{\underline{\lambda}_1 \cdot \underline{\phi}_1(\underline{x}) + \underline{\lambda}_2 \cdot \underline{\phi}_2(\underline{x})}$

//

"Searching over Statistics" is complicated and difficult

∴ See papers by Della Pietra, Della Pietra, Lafferty

Zhu, Wu, Mumford.

Beyond scope of this course.