Page 1) Notes to Clarify Material on Tuesday Afternoon

Note Title 6/28/2011

Exponential Model $P(d|\underline{\lambda}) = \dfrac{e^{\underline{\lambda}\cdot\underline{\phi}(d)}}{Z[\underline{\lambda}]}$

Data $D = \{d^\mu : \mu = 1 \text{ to } N\}$,
$P(D|\underline{\lambda}) = \prod_\mu P(d^\mu|\underline{\lambda})$

Maximum Likelihood (ML) $\hat{\underline{\lambda}} = \text{ARG MAX } P(D|\underline{\lambda})$

reduces to solve for $\hat{\underline{\lambda}}$: $\frac{1}{N}\sum_{\mu=1}^{N}\underline{\phi}(d^\mu) = \sum_{\underline{d}} \underline{\phi}(d)\, P(\underline{d}|\hat{\underline{\lambda}})$

Probability of the data with best $\underline{\lambda}$.

$$P(D|\hat{\underline{\lambda}}) = \prod_\mu P(d^\mu|\hat{\underline{\lambda}})$$

Intuitively, if $P(D|\hat{\underline{\lambda}})$ is big - i.e. the data is very probable - then we think that the model fits the data well.

Model Selection:

Suppose we have two models

$$P_1(d|\lambda_1) = \frac{1}{Z_1[\lambda_1]} e^{\underline{\lambda}_1\cdot\underline{\phi}_1(d)}, \quad P_2(d|\lambda_2) = \frac{1}{Z_2[\lambda_2]} e^{\underline{\lambda}_2\cdot\underline{\phi}_2(d)}$$

with different statistics.

Which model is best?

For each model, find $\hat{\underline{\lambda}}_i = \text{ARG MAX}_{\underline{\lambda}_i} P_i(D|\underline{\lambda}_i)$

the best parameter for each model.

Evaluate: $P_1(D|\hat{\lambda}_1) = \prod_{\mu=1}^{N} P_1(d_\mu|\hat{\lambda}_1)$

and $P_2(D|\hat{\lambda}_2) = \prod_{\mu=1}^{N} P_2(d_\mu|\hat{\lambda}_2)$

Select model 1, if $P_1(D|\hat{\lambda}_1) > P_2(D|\hat{\lambda}_2)$

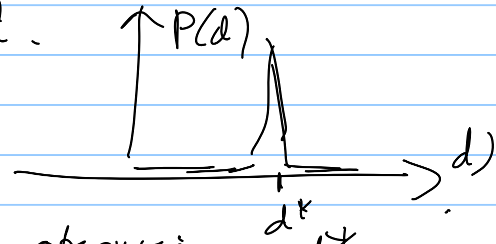
model 2, if $P_2(D|\hat{\lambda}_2) > P_1(D|\hat{\lambda}_1)$

Model Selection (type I)

Note — there is an interpretation of this based on entropy.

Entropy $H[P] = -\sum_{d} P(d)\log P(d)$

Entropy is a measure of how much information we gain from making an observation d.

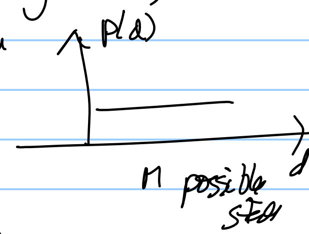


Suppose: $P_0(d) = \delta(d - d_*)$
$= 0$, if $d \neq d_*$

We get no information from observing d^*, because we know that it is the on observation we can get.

$H(P_0) = 0$ $\quad (0\log 0 = 0,\ 1\log 1 = 0)$

Suppose: $P_1(d) = U(d)$ ← uniform distibution



then we get $H(P_1) = \log M$ ← no. of possible values of d

Result.

$$\log P(D|\hat{\lambda}) = \sum_{\mu} \log P(d_\mu|\hat{\lambda})$$

$$= \sum_{\mu} \hat{\lambda}\cdot\phi(d_\mu) - N\log Z[\hat{\lambda}]$$

ML estimate

Entropy of $P(d|\hat{\lambda})$

$$= -\sum_{d} P(d|\hat{\lambda})\log P(d|\hat{\lambda})$$

$$= -\sum_{d} P(d|\hat{\lambda})\{\hat{\lambda}\cdot\phi(d) - \log Z[\hat{\lambda}]\},$$

But, $\frac{1}{N}\sum_{\mu}\phi(d_\mu) = \sum_{d}\phi(d)P(d|\hat{\lambda})$ definition of ML

Hence

$$\log P(D|\hat{\lambda}) = -N\,\text{Entropy}(P(d|\hat{\lambda}))$$

$$P(D|\hat{\lambda}) = e^{-N\,\text{Entropy}(P(d|\hat{\lambda}))}$$

So $P(D|\hat{\lambda})$ is big if Entropy $P(d|\hat{\lambda})$ is small.

Maximum Likelihood corresponds to minimizing Entropy

The best model to fit data has lowest energy, hence best ability to predict.

Feature Pursuit:

Make a dictionary $A = \{\phi_1(\,), \ldots, \phi_N(\,)\}$, of possible features.

Task: want to construct a probability model.

$$P(d \mid \{\lambda_i\}) = \frac{1}{Z[\{\lambda_i\}]} e^{\sum_{i=1}^{N} \lambda_i \phi_i(d)}$$

to model the data,

Want to keep model simple — use only a few of the features (also data limitations — later in course).

want $\lambda_i = 0$ for most i

Two tasks:

(i) selection — which features to use (i.e. to have $\lambda_i \neq 0$)

(ii) weighting — how to weight features and assign λ_i ?

This is a hard search problem (easier for discriminative learning — later in the course).

Strategy: feature Pursuit. → Della Pietra^2, Lafferty → Zhu, Wu, Mumford.

(1) Find best model with one feature only

Calculate: $\hat{i}$ s.t. $P_{\hat{i}}(D \mid \hat{\lambda}_{\hat{i}}) \geq P_i(D \mid \hat{\lambda}_i)$ for all $i = 1$ to N

Here $P_i(d \mid \lambda_i) = \frac{1}{Z[\lambda_i]} e^{\lambda_i \phi_i(d)}$, $\hat{\lambda}_i$ by ML.

This selects feature $\phi_{\hat{i}}$ and assigns it weight $\hat{\lambda}_{\hat{i}}$

(2) Next add another feature/statics:

Consider all models of form:

$$P_{\hat{i}j}(d \mid \hat{\lambda}_{\hat{i}}, \lambda_j) = \frac{1}{Z[\hat{\lambda}_{\hat{i}}, \lambda_j]} e^{\underbrace{\hat{\lambda}_{\hat{i}} \cdot \phi_{\hat{i}}(d)}_{\text{feature selected already}} + \underbrace{\lambda_j \cdot \phi_j(d)}_{\text{new feat.}}}$$

Select the second feature $\hat{j}$ by finding

$$P_{\hat{i}\hat{j}}(D|\underline{\hat{\lambda}}_{\hat{i}}, \underline{\hat{\lambda}}_{\hat{j}}) \geq P_{\hat{i}j}(D|\underline{\hat{\lambda}}_{\hat{i}}, \underline{\hat{\lambda}}_{j})$$

for all $j \neq \hat{i}$

Proceed to select and weight the third, fourth, fifth, ... features and weight them.

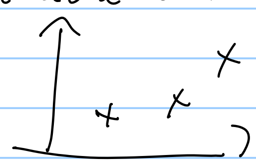
When to stop?

Adding a new feature allows the model to fit the data better

→ ie. $P_{\hat{i}\hat{j}}(D|\underline{\hat{\lambda}}_{\hat{i}}, \underline{\hat{\lambda}}_{\hat{j}}) \geq P_{\hat{i}}(D|\underline{\hat{\lambda}}_{\hat{i}})$

(because the model with two features has more flexibility and can find the data better

— example



fit data to line
$y = a + bx$
or to curve
$y = a + bx + cx^2$.
easier, more flexibility

So, stop if increase by adding a new feature falls below a threshold:

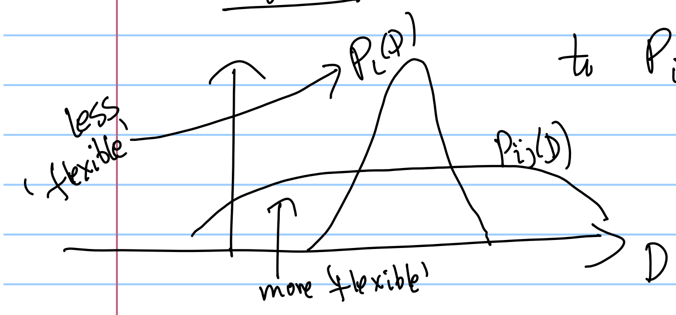
→ ie. stop with $P_{\hat{i}}(D|\underline{\hat{\lambda}}_{\hat{i}})$

if $P_{\hat{i}\hat{j}}(D|\underline{\hat{\lambda}}_{\hat{i}}, \underline{\hat{\lambda}}_{\hat{j}}) \leq P_{\hat{i}}(D|\underline{\hat{\lambda}}_{\hat{i}}) + T$ ← threshold.

Note: A more advanced form of model selection will avoid this. → Occam's Razor (often impractical)

Requires: Comparing $P_i(D) = \sum_{\lambda_i} P_i(D|\lambda_i)$

to $P_{ij}(D) = \sum_{\lambda_i, \lambda_j} P_{ij}(D|\lambda_i, \lambda_j)$



$\sum_D P_{ij}(D) = 1$

$\sum_D P_i(D) = 1)$

Expectation-Maximization:

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

observed ↗ ↖ hidden

Do ML on $P(\underline{d}|\underline{\lambda}) = \sum_{\underline{h}} P(\underline{d}, \underline{h}|\underline{\lambda})$ to estimate $\underline{\lambda}$

Minimize $-\log P(\underline{d}|\underline{\lambda})$ with respect to $\underline{\lambda}$

Add a new variable $q(\underline{h})$, a distribution over the hidden variables $\sum_{\underline{h}} q(\underline{h}) = 1$

Now $P(\underline{h}|\underline{d}, \underline{\lambda})$ is the probability of the hidden variable $\underline{h}$, if we know the parameter $\underline{\lambda}$.

Formally $P(\underline{h}|\underline{d}, \underline{\lambda}) = \dfrac{P(\underline{h}, \underline{d}|\underline{\lambda})}{\sum_{\underline{h}} P(\underline{h}, \underline{d}|\underline{\lambda})}$

(calculating $\sum_{\underline{h}} P(\underline{h}, \underline{d}|\underline{\lambda})$ may be difficult — see later.)

Want $q(\underline{h})$ to be close to $P(\underline{h}|\underline{d}, \underline{\lambda})$

Kullback-Leibler $\sum_{\underline{h}} q(\underline{h}) \log \dfrac{q(\underline{h})}{P(\underline{h}|\underline{d}, \underline{\lambda})} \geq 0$

$= 0$ if $q(\underline{h}) = P(\underline{h}|\underline{d}, \underline{\lambda})$

Define:

$$F[\underline{\lambda}, q(.)] = -\log P(d|\underline{\lambda}) + \sum_{\underline{h}} q(\underline{h}) \log \frac{q(\underline{h})}{P(\underline{h}|\underline{d}, \underline{\lambda})}$$

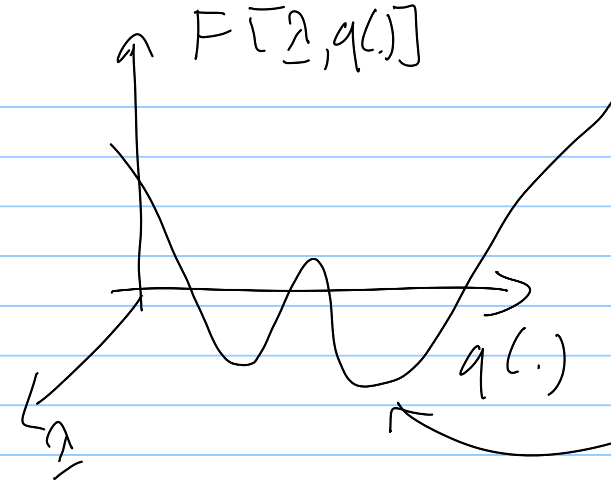
This is a function of $\underline{\lambda}$ and $q(.)$.

Its global minimum occurs at.

$\hat{\underline{\lambda}} = \text{ARG MIN}_{\underline{\lambda}} (-\log P(\underline{d}|\underline{\lambda}))$, ML. $\underline{\lambda}$

and at $\hat{q}(\underline{h}) = P(\underline{h}|\underline{d}, \hat{\underline{\lambda}})$ — makes second term 0.

Page 6.



This is non-convex. Will have several minima in general.

Global minimum at $\hat{\underline{\lambda}}, \hat{q}(.)$

Minimize by coordinate descent.

(1) At state $\underline{\lambda}^t$,

Solve for $q^t(h) = \arg\min_{q(\,)} F[\underline{\lambda}^t, q(\,)]$

Solution: $q^t(h) = P(h|d, \underline{\lambda}^t)$

(requires computing $\frac{P(h,d|\underline{\lambda}^t)}{\sum_h P(h,d|\underline{\lambda}^t)}$)

(2) At state $q^t(h)$

Compute $\underline{\lambda}^{t+1} = \arg\min_{\underline{\lambda}} F[\underline{\lambda}, q^t]$

Solution.

$$\sum_h q^t(h)\, \underline{\phi}(\underline{d}, \underline{h}) = \sum_{h,d} \underline{\phi}(\underline{d}, \underline{h}) P(\underline{h}, \underline{d}|\underline{\lambda})$$

data statistics and expected $\underline{h}$ w.r.t. $q^t(h)$

model statistics

Repeat Steps.

Performance depends on the initial condition $\underline{\lambda}^{t=0}$.

Previous notes (Tuesday) gave results for the extended case when we have data $D = \{\underline{d}_\mu : \mu = 1 \text{ to } N\}$

Then replace $F[\underline{\lambda}, q(.)]$ by

$$-\sum_\mu \log P(d_\mu|\lambda) + \sum_\mu \sum_{h_\mu} q_\mu(h_\mu) \log \frac{q_\mu(h_\mu)}{P(h_\mu|d_\mu, \lambda)}$$