Learning

- Being able to learn is of critical importance
  - Need to refine our uncertain knowledge using available data

- Almost all methods of learning require some form of inference over variables.

- Sometimes very approximate inference can still provide sufficient “signal” for learning.

- This week: learning is just another form of inference. It is inference over parameters.
  - Inference: inference using information from data items.
  - Learning: inference using information from data sets.
Previous lecture:

But where do the numbers come from?
If we don’t know the numbers what should we do?
Parameters

- If we don’t know the numbers we can represent them using parameters.

- But parameters are just unknown items with uncertainty.

- How are they different from random variables?
Parameters

- How are they different from random variables?

- They are not different
  - In fact we can just include parameters in our model and infer them in the same way.

- They are different
  - Parameters are extrinsic rather than intrinsic quantities. (I’ll Explain)
  - To infer them we need to condition on data sets not data items.
Parameterising a Distribution.

- Previously

\[ P(x) = \frac{1}{Z} \exp(-E(x)) \]

- Now

\[ P(D|\theta) = \frac{1}{Z(\theta)^{N_D}} \prod_{n=1}^{N_D} \exp(-E(x^n|\theta)) \]

where \( D \) denotes the data set \( D = \{x_1, x_2, \ldots, x^{N_D}\} \).
\( \theta \) denotes the collection of all the parameters.

- Note same parameter for each \( n \).
Using Parameters

For $P(\text{Toothache}, \text{Cavity})$ we can write

<table>
<thead>
<tr>
<th></th>
<th>Toothache = true</th>
<th>Toothache = false</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cavity = true</td>
<td>0.04</td>
<td>0.06</td>
</tr>
<tr>
<td>Cavity = false</td>
<td>0.01</td>
<td>0.89</td>
</tr>
</tbody>
</table>

For $P(\text{Toothache}, \text{Cavity})$ we can write

<table>
<thead>
<tr>
<th></th>
<th>Toothache = true</th>
<th>Toothache = false</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cavity = true</td>
<td>$\theta_1$</td>
<td>$\theta_3$</td>
</tr>
<tr>
<td>Cavity = false</td>
<td>$\theta_2$</td>
<td>$1 - \theta_1 - \theta_2 - \theta_3$</td>
</tr>
</tbody>
</table>
Plate Notation

- Represent repetition in graphical models as a ‘plate’:

- Items inside the plate, or crossing the plate are repeated each time.
- Items outside are not repeated.

$n=1,2,...,N$
\[ P(\theta) \prod_{n=1}^{3} P(x_n)P(y_n|x_n)P(z_n|x_n, \theta) = \]
Learning as Inference

- Given data $D = \{v^n | n = 1, 2, ..., N_D\} = v^{1:n}$. Have full model
  $$ \prod_{n=1}^{N_D} P(v^n, x^n, h^n | \theta) $$

- Define a prior distribution $P(\theta)$ over parameters to give joint model
  $$ P(v^{1:n}, x^{1:n}, h^{1:n}, \theta) = \left[ \prod_{n=1}^{N_D} P(v^n, x^n, h^n | \theta) \right] P(\theta) $$

- Now consider a new test case (not in the training set). Index by $n = 0$
  $$ P(v^{0:n}, x^{0:n}, h^{0:n}, \theta) = \left[ \prod_{n=0}^{N_D} P(v^n, x^n, h^n | \theta) \right] P(\theta) $$

- Compute what we want via inference:
  $$ P(x^0 | v^0, D) \propto P(x^0, v^0 | D) = \sum_{\theta, h^{0:n}, x^{1:n}} \left[ \prod_{n=0}^{N_D} P(v^n, x^n, h^n | \theta) \right] P(\theta) $$
Prior

- Note we needed prior $P(\theta)$ to do this.
- Prior comes from assumptions about parameters.
- Sometimes people choose reasonable assumptions.
- Sometimes people choose assumptions that make the maths easy.
- Sometimes people choose a vague prior (but it is still an assumption)
- Note also that the “sum” over $\theta$ is really likely to be an integral over $\theta$ as parameters will likely be real valued.
- We will be doing some multivariate integrals...
Independent Data Items

- If, conditioned on the parameters, each data item is independent.
  - Can use plate notation.
  - The test data is only connected to the training data via the parameters.

\[ P(x^0, v^0|D) = \int d\theta P(x^0, v^0|\theta)P(\theta|D) \]

- \( P(\theta|D) \) is the *posterior* distribution.
- It is what we now know about the parameters having seen the data.
Example

$p$ denotes the probability of a 1 turning up. What is the probability that the next item $x^*$ is 1?

$p$ takes one value for all data items. But we don’t know what that is! Use a distribution to represent and compute with the uncertainty.

\[\int dp \, P(x^* = 1|p)P(p|D)\text{ where } P(p|D) = \frac{P(D|p)P(p)}{P(D)}\]

Now $P(D|p) = \prod_n P(x^n|p) = p^{N_1}(1-p)^{N_0} = p^9(1-p)^{11}$

so $P(p|D) = \frac{23!}{10!12!}p^{10}(1-p)^{12}$

meaning $P(x^* = 1) = \int dp \, P(x^* = 1|p)P(p|D)$. Computing this gives $11/24$. (Hint: look up Beta distribution).

Note $N_1 = 9$ is the number of ones, and $N_0 = 11$ is the number of zeros.
Learning can be hard.

- Why can learning be hard?

\[ P(v, h|\theta) = \frac{1}{Z(\theta)} \exp(-E(v, h; \theta)) \]

where

\[ Z(\theta) = \int dv dh \exp(-E(v, h; \theta)) \]

- \( Z \) depends on the parameters.
- It is not always easy to compute.
Our Journey

- Introduction to Learning
- Next: Q&A
- Next: Learning exponential family models, with examples.