What makes a good tutorial?

Iain Murray

School of Informatics, University of Edinburgh

What do the students want from their tutorials?

Answers But I give them *detailed* written answers afterwards...

Another point of view

Contact hour in small group

To pass the exam

To learn about the topic

To gain skills?

What are tutorials most useful for?

What I value

Communication skills

Critical thinking

Becoming confident of answers

Feedback

Discussion? See also *workshops*, *labs*, *pals*, ...

Some exercises / answers

to discuss

As part of a derivation, we may need to identify the probability density function of a vector up to a constant. For example:

$$p(\mathbf{x}) \propto \exp\left(-\mathbf{x}^\top A\mathbf{x} - \mathbf{x}^\top \mathbf{c}\right),$$

where A is a symmetric invertible matrix. As this distribution is proportional to the exponential of a quadratic in $\mathbf{x}$, it is a Gaussian: $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \Sigma)$.

Identify which Gaussian $\mathbf{x}$ comes from by identifying the mean $\boldsymbol{\mu}$ and covariance Σ in terms of A and $\mathbf{c}$. The easiest method is to compare $p(\mathbf{x})$ to the standard form for the multivariate Gaussian PDF (given in class).

The answer you should be able to show is:

$$\Sigma = \frac{1}{2}A^{-1}, \quad \boldsymbol{\mu} = -\frac{1}{2}A^{-1}\mathbf{c}.$$

$$\begin{aligned}
\log \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \Sigma) &= -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}) + \text{const.} \\
&= -\frac{1}{2}\mathbf{x}^\top \Sigma^{-1}\mathbf{x} + \frac{1}{2}\mathbf{x}^\top \Sigma^{-1}\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{\mu}^\top \Sigma^{-1}\mathbf{x} + \text{const.} \\
&= -\frac{1}{2}\mathbf{x}^\top \Sigma^{-1}\mathbf{x} + \mathbf{x}^\top \Sigma^{-1}\boldsymbol{\mu} + \text{const.} \quad (\Sigma \text{ and } \Sigma^{-1} \text{ are symmetric})
\end{aligned}$$

In comparison the quadratic form given in the question is:

$$\log p(\mathbf{x}) = -\mathbf{x}^\top A\mathbf{x} - \mathbf{x}^\top \mathbf{c} + \text{const.}$$

We can read off the mean and covariance by comparing the coefficients of these forms. Comparing the quadratic term:

$$-\frac{1}{2}\Sigma^{-1} = -A \quad \Rightarrow \quad \boxed{\Sigma = \tfrac{1}{2}A^{-1}}.$$

Comparing the linear term:

$$\Sigma^{-1}\boldsymbol{\mu} = -\mathbf{c} \quad \Rightarrow \quad \boxed{\boldsymbol{\mu} = -\Sigma\mathbf{c} = -\tfrac{1}{2}A^{-1}\mathbf{c}}.$$

b) A common programming mistake is to forget the minus sign in either the descent procedure or in the gradient evaluation. As a result one unintentionally writes a procedure that does $\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^{(t)} + \eta \nabla_{\mathbf{w}} E$. What happens?

b) Getting the sign of the gradient wrong means the algorithm maximizes E rather than minimizing it. Often when we're minimizing a cost, it has no finite maximum. For example in linear regression we can make the squared error as large as we like by setting extreme weights. So typically $E(\mathbf{w})$ and the magnitude of the weights increases towards infinity over time. In practice the code will often crash due to a numerical error.

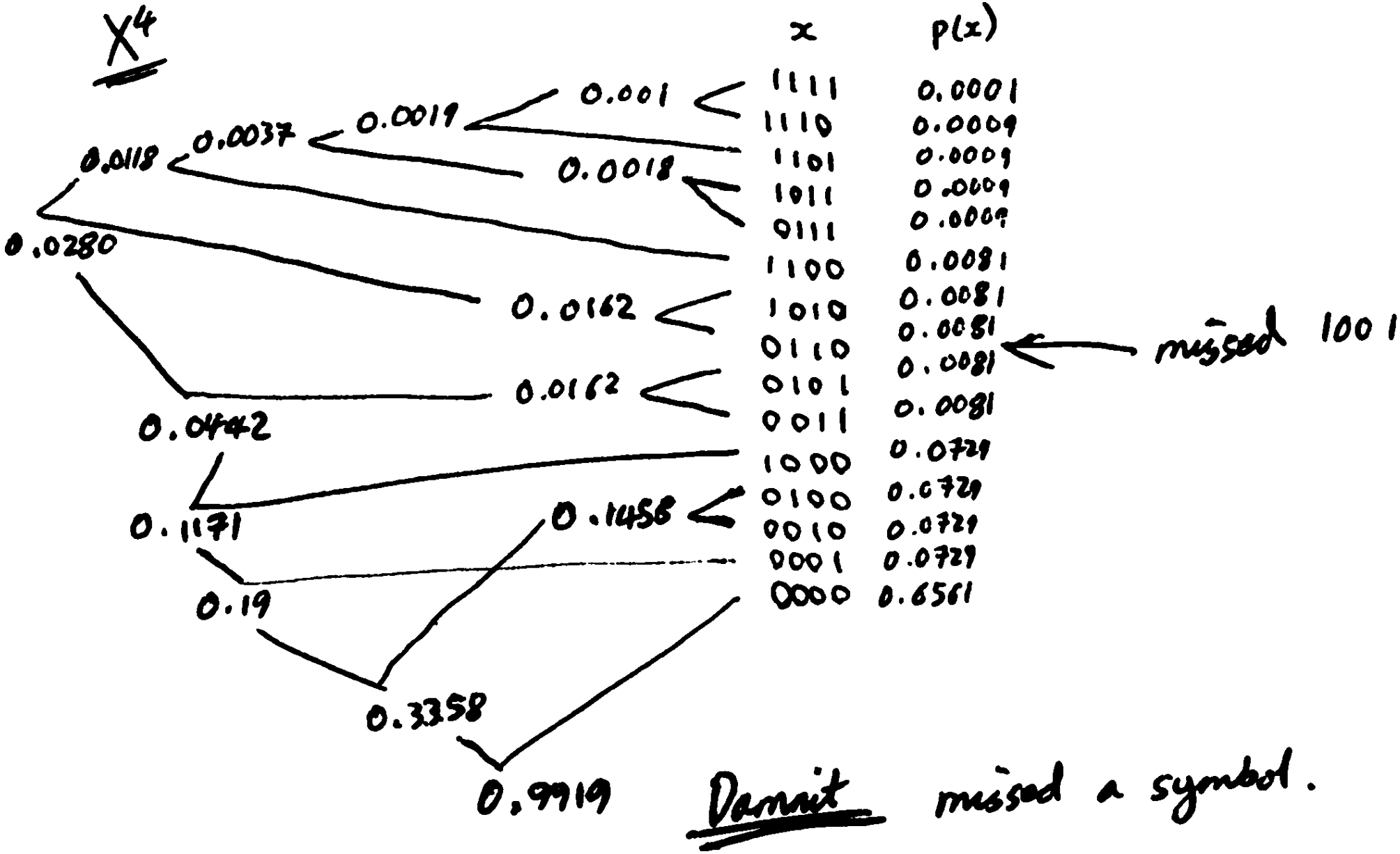
We have a dataset of inputs and outputs $\{\mathbf{x}^{(n)}, y^{(n)}\}_{n=1}^{N}$, describing N preparations of cells from some lab experiments. The output of interest, $y^{(n)}$, is the fraction of cells that are alive in preparation n. The first input feature of each preparation indicates whether the cells were created in lab `A`, `B`, or `C`. That is, $x_1^{(n)} \in \{\texttt{A}, \texttt{B}, \texttt{C}\}$. The other features are real numbers describing experimental conditions such as temperature and concentrations of chemicals and nutrients.

a) Describe how you might represent the first input feature and the output when learning a regression model to predict the fraction of alive cells in future preparations from these labs. Explain your reasoning.

...

c) There's a debate in the lab about how to represent the other input features: log-temperature or temperature, and temperature in Fahrenheit, Celsius or Kelvin? Also whether to use log concentration or concentration as inputs to the regression. Discuss ways in which these issues could be resolved.

   Harder: there is a debate between two different representations of the output. Describe how this debate could be resolved.


X^4
x
p(x)
1111 0.0001
1110 0.0009
1101 0.0009
1011 0.0009
0111 0.0009
1100 0.0081
1010 0.0081
0110 0.0081
0101 0.0081
0011 0.0081
1000 0.0729
0100 0.0729
0010 0.0729
0001 0.0729
0000 0.6561
0.001
0.0019
0.0037
0.0118
0.0018
0.0280
0.0162
0.0162
0.04442
0.1458
0.1171
0.19
0.3358
0.9919
missed 1001
Damnit missed a symbol.

...and finally:

$$\boxed{p_1^* = \frac{1}{(1-f)(1+2^{H_2(f)/(1-f)})},}$$

which is equivalent to the previous answer.

Are these messy expressions correct? It is important to treat results with suspicion. To be honest, the first thing I did was check my first expression against the answer in the book numerically. But we won't always have answers in books to check. (And who says they got it right?) In general, extreme cases are often easy to check. Here, the channel may reduce to one that is easier to analyse.

When there is no noise, $f = 0$, we are just source coding: we want to use the symbols equally often, and indeed we recover $p_1^* = 1/2$ in this case.

As $f \to 1$, the channel only ever emits a single symbol, so the capacity will tend to zero. When that happens, it doesn't matter what input distribution we use, the mutual information is zero no matter what. For interest, we can see what input distribution we should use as we approach this limit:

$$\lim_{f \to 1} p_1^* = \lim_{f \to 1} f^{f/(1-f)} = \frac{1}{-}.$$

...

Closing thoughts

Set expectations

Finding what works is iterative

taking >1 group useful!

Ask what are main goals of this tutorial?

Give your lecturer feedback