David L. Dowe [© Jun. 2011]

(www.csse.monash.edu.au/∼dld ;

david dot dowe At monash dot edu)

MML analysis of *all* data-sets and much more

(including theories of intelligence and automating database normalisation)

Statistical invariance, and
Statistical consistency

My papers (Dowe & Wallace, 1998; Comley & Dowe, 2003, 2005) first to show how to use both discrete (multi-state, categorical) and continuous valued variables in MML Bayesian nets.

Desiderata (in inference)

Statistical invariance

– Circle: $\hat{A} = \pi\ \hat{r}^2$

– Cube: $\hat{l} \ = \ \hat{A}^{1/2} \ = \ \hat{V}^{1/3}$

– Cartesian/Polar: $(\hat{x}, \hat{y}) = (\hat{r}\cos(\hat{\theta}), \hat{r}\sin(\hat{\theta}))$

Statistical consistency

As we get more and more data, we converge more and more closely to the true underlying model

(But what if data-generating source is outside our model space?)

Efficiency

Not only are we statistically consistent, but as we get more and more data we converge as rapidly as is possible to any underlying model.

Some methods of inference

Maximum Likelihood: Given data D, choose (probabilistic) hypothesis H to maximise $f(D|H)$ and minimise $-\log f(D|H)$.

– Statistically invariant — but tends to over-fit, "finding" non-existent patterns in random noise

– Also, how do we choose between models of increasing complexity and increasingly good fit e.g., constant, linear, quadratic, cubic, ...?

– Also, maximum likelihood chooses the hypothesis to make the already observed data as likely as possible.

But, shouldn't we choose H so as to maximise $Pr(H|D)$?

Bayesianism, prior prob's, $Pr(H|D)$
Prior probability, $Pr(H)$

$Pr(H).Pr(D|H) = Pr(H\&D) =$
$Pr(D\&H) = Pr(D).Pr(H|D)$

So, $Pr(H|D) = \frac{Pr(H).Pr(D|H)}{Pr(D)} =$
$\frac{1}{Pr(D)}(Pr(H).Pr(D|H))$

$$posterior(H|D) = \frac{prior(H)\ .\ likelihood(D|H)}{marginal(D)}$$

Probability vs probability *density*

What is your (friend's) height? weight? *Measurement accuracy* - used in MML in lower bound for some parameter estimates, but overlooked and ignored in classical approaches

Information Theory

$$max_H \; Pr(H|D) =$$
$$max_H \frac{1}{Pr(D)}(Pr(H).Pr(D|H)) =$$
$$max_H \; Pr(H).Pr(D|H) =$$
$$min_H \; -\log Pr(H) \; -\log Pr(D|H)$$

Can do this if everything is a probability and not a density, whereupon $l_i = -\log_2 p_i$ is the binary code-length of an event of prob' p_i

$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{21}$
$\frac{1}{4}$	$\frac{1}{4}$	$\frac{2}{21}$
$\frac{1}{4}$	$\frac{1}{4}$	$\frac{3}{21}$
$\frac{1}{8}$	$\frac{1}{4}$	$\frac{6}{21}$
$\frac{1}{16}$		$\frac{4}{21}$
$\frac{1}{16}$		$\frac{5}{21}$

Uniqueness result [Dowe (2008a-b, 2011)] that logarithm-loss is unique invariant "*true*" scoring system.