Limitation of MLE

- Over-reliance on data. If data is limited, estimates can be very wrong. Example, multinomial probabilities could be zero for categories that do not appear in a limited sample.

- No indication on the uncertainty of the estimated parameters. Example, for a Bernoulli parameters whether estimation is made from two with 50% heads or 1000 examples with 50% heads, the estimated parameter is the same.

- No mechanism to specify human's prior knowledge of the parameters.

Bayesian Estimation

$X \sim \text{Bernoulli}(x; w)$ $\boxed{w}$ = probability of head.

- Treat the parameters as a random variable. Human's specify their prior probability knowledge of the values of the parameters as a distribution $p(\mathbf{w})$

Common prior for Bernoulli: $p(w) \sim \text{Uniform}([0,1])$

- Given a dataset D, we express the probability of the data in terms of the parameter as:

$$P(D|w) = L(w) = \prod_{i=1}^{N} f(x^i; w)$$ [Used in MLE)

- During parameter estimation, apply Bayes rule to estimate probability of parameters given data and prior.

$$p(w|D) = \frac{p(D|w)\,p(w)}{\int_w p(D|w)\,p(w)}$$

← In general difficult.

← posterior distribution.

Deployment

Exact method: $f(x|\mathbf{w}, D) = \int_{\mathbf{w}} p(\mathbf{w}|D) f(x|\mathbf{w}) dw$

Complicated in general, but for simple priors and distributions it gives rise to helpful estimates.

Examples:

Bayesian estimate of paramters for common distributions

(1) Bernoulli: $\frac{n_1(D)+1}{N+2}$ when prior $p(w) \sim \text{Uniform}([0,1])$

(2) Multinominal distribution: $p_j = \frac{n_j(D)+1}{N+k}$ if $p(p_1, p_2, \ldots p_k) \sim \text{Dirichlet}(; [1, 1 \ldots 1])$

[not in syllabus]