

1/26

MLE: Application to broader problems. ex. Observe samples of X

$D: d[1] \dots d[M]$ where $d[i]$ is in $\{x_1 \dots x_N\}$

$p(d; \theta)$ – assigns probability to data

$p(d)$ given θ , should have legal distribution (ie > 0 , sums to 1)

Define parameter space to get a legal distribution.

Ex $P(\text{thumbtack})(x; \theta) = \theta$ for $x=H$, $1-\theta$ for $x=T$

defining likelihood function: probability model assigns to training data

$L(\theta; D) = P(d[1] \dots d[M]; \theta) = \prod_{m=1}^M (P(d[m]; \theta))$

Assume samples are i.i.d.

Use MLE to find parameter values. Given data set D it chooses parameters $\hat{\theta}$ such that $L(\hat{\theta}; D) = \max_{\theta} L(\theta; D)$

Simple Bayesian Network

$X \rightarrow Y$ $\text{Val}(X) = \{x_0, x_1\}$ $\text{Val}(Y) = \{y_0, y_1\}$

What are the parameters of this network? $p(X, Y) = p(X)p(Y|X)$ use CPD table.

M training samples that are two-ples: $\langle x[m], y[m] \rangle$

$L(\theta; D)$ (under i.i.d. Assumption) = $\prod_{m=1}^M p(x[m], y[m]; \theta) = \prod (p(x[m]; \theta_x) p(y[m]|x[m]; \theta_{Y|X}))$

L is decomposed into two terms.

$\hat{\theta}(x)$ is the θ that maximizes $L(\theta; D)$. The second term does not depend on $\theta(x)$ and so we only need to maximize the first term. $\hat{\theta}(Y|X)$ is the θ that maximizes the second term.

We can similarly decompose the second term.