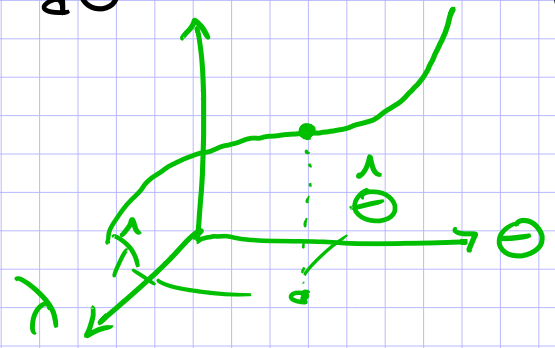
Lecture 5: 2/1/17

Aside: Method of Lagrange Multipliers

① Given problem: $\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ f(\theta)$ s.t. $g(\theta) = c$

② Introduce Lagrangian: $L(\theta, \lambda) = f(\theta) - \lambda(g(\theta) - c)$

③ Solve $\frac{dL(\theta,\lambda)}{d\theta} = 0$ and $\frac{dL(\theta,\lambda)}{d\lambda} = 0$



Back to MLE of Cat.

✡ Apply the Method Lagrange Mult.

$$L(\vec{\theta}, \lambda) = \ell(\theta) - \lambda\left(\sum_{k=1}^{K}\theta_k - 1\right)$$

$$= \sum_{k=1}^{K} N_k \log\theta_k - \lambda\left(\sum_{k=1}^{K}\theta_k - 1\right)$$

N_k: # times that k appears in D

$$\frac{dL(\vec{\theta},\lambda)}{d\theta_j} = \frac{N_j}{\theta_j} - \lambda = 0 \Rightarrow \theta_j = \frac{N_j}{\lambda}$$

$$\frac{dL(\vec{\theta},\lambda)}{d\lambda} = \sum_{k=1}^{K}\theta_k - 1 = 0 \Rightarrow \sum_{k=1}^{K}\theta_k = 1$$

$\forall j$

$$\theta_j = \frac{N_j}{\sum_{k=1}^{K} N_k}$$

Apply MLE to Data

Ex #1: $D = \{x^{(1)}, \ldots, x^{(N)}\} = [1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 3]$

$x^{(i)} \sim \text{Categorical}(\vec{\theta}) \quad K \triangleq |\vec{\theta}| = 3$

$$\vec{\theta}_{MLE} = \begin{bmatrix} \theta_{1,MLE} \\ \theta_{2,MLE} \\ \theta_{3,MLE} \end{bmatrix} = ? = \begin{bmatrix} 6/11 \\ 4/11 \\ 1/11 \end{bmatrix}$$

Ex #2: Same dataset D as above.
But $K = 6$

$$\vec{\theta}_{MLE} = \begin{bmatrix} \theta_{1,MLE} \\ \theta_{2,MLE} \\ \vdots \\ \theta_{6,MLE} \end{bmatrix} = ? = \begin{bmatrix} 6/11 \\ 4/11 \\ 1/11 \\ 0/11 \\ 0/11 \\ 0/11 \end{bmatrix}$$

Suppose D is from six-sided die.

Is MLE still a good estimate?

Learning from Data (Bayesian)

MAP Estimation:

- "Maximum a posteriori" $\Longleftrightarrow$ MAP
- Bayesian statistics
- key idea: inject world knowledge into estimation problem via a prior over parameters
- MAP estimate is the mode of the posterior

Def: posterior

$$\underbrace{p(\theta|D)}_{\text{posterior}} = \frac{\overbrace{p(D|\theta)}^{\text{likelihood}}\overbrace{p(\theta)}^{\text{prior}}}{p(D)}$$

Bayes' Rule

Opt. for MAP Est.

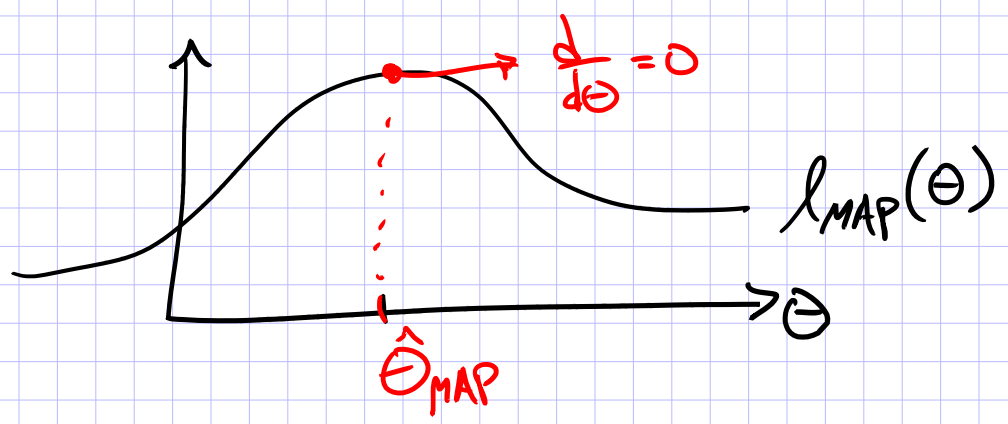
$$\begin{aligned}
\hat{\theta}_{MAP} &= \underset{\theta\in\Theta}{\operatorname{argmax}}\ p(\theta|D) \\
&= \underset{\theta\in\Theta}{\operatorname{argmax}}\ \log p(\theta|D) \\
&= \underset{\theta\in\Theta}{\operatorname{argmax}}\ \log\left[\frac{p(D|\theta)p(\theta)}{p(D)}\right] \\
&= \underset{\theta\in\Theta}{\operatorname{argmax}}\ \log[p(D|\theta)p(\theta)] \\
&= \underset{\theta\in\Theta}{\operatorname{argmax}}\ \ell_{MAP}(\theta)
\end{aligned}$$

where $\ell_{MAP}(\theta) \triangleq$

$p(D)$ wrt. θ is constant

why convenient?

$$p(D) = \int_\theta p(D|\theta)p(\theta)\,d\theta$$



Ex: MAP Estimate of a Bernoulli

① Write the likelihood

$$X^{(i)} \sim \text{Bernoulli}(\phi) \quad \text{where} \quad 0 \le \phi \le 1$$

$$p(X=x) = \begin{cases} \phi & \text{if } x=1 \\ 1-\phi & \text{if } x=0 \end{cases}$$

$$= (\phi)^x (1-\phi)^{(1-x)} \quad \text{notation trick}$$

$$\Rightarrow p(D|\phi) = \prod_{i=1}^{N} (\phi)^{x^{(i)}} (1-\phi)^{(1-x^{(i)})} = \phi^{N_1} (1-\phi)^{N_0}$$

where $N_1 = \#\,1\text{s}$

$N_0 = \#\,0\text{s}$

② Write prior

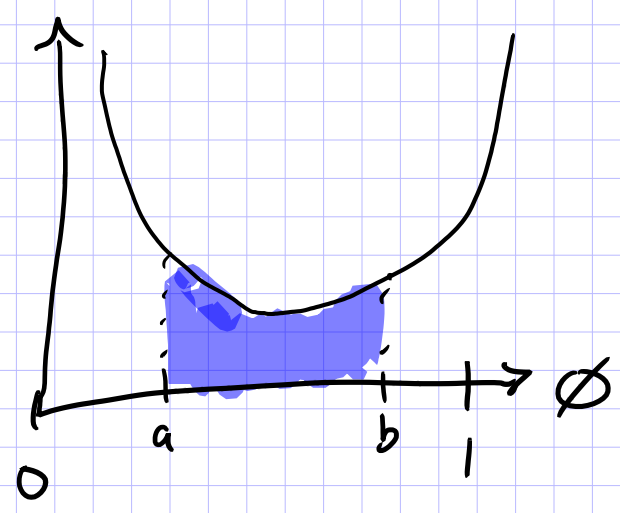
$$p(\phi) = \begin{cases} ? & \text{if } 0 \le \phi \le 1 \\ 0 & \text{otherwise} \end{cases}$$

$$p(\phi) \sim \text{Beta}(\alpha, \beta)$$

probability density function:

$$f(\phi|\alpha,\beta) = \frac{1}{B(\alpha,\beta)} \phi^{(\alpha-1)} (1-\phi)^{(\beta-1)}$$

$$p(a \le \phi \le b) = \int_a^b f(\phi|\alpha,\beta)\, d\phi$$



(3) Write Joint (likelihood times prior)

$$p(D|\theta)p(\theta) = [\phi^{N_1}(1-\phi)^{N_0}][\phi^{(\alpha-1)}(1-\phi)^{(\beta-1)}] \cdot \frac{1}{B(\alpha,\beta)}$$

$$= \underbrace{\phi^{(N_1+\alpha-1)}(1-\phi)^{(N_0+\beta-1)}} \cdot \frac{1}{B(\alpha,\beta)}$$

identical to likelihood where $(N_1+\alpha-1)$ "heads" and $(N_0+\beta-1)$ "tails"

$$D = [\underbrace{000\cdots0}_{N_0}\ \underbrace{00\cdots0}_{\beta-1}\ \underbrace{1111\cdots1}_{N_1}\ \underbrace{11\cdots1}_{\alpha-1}]$$

(4) Solve the optimization

$$\phi_{MAP} = \underset{\phi \in \mathbb{R}}{\operatorname{argmax}}\ \ell_{MAP}(\phi) \quad \text{s.t. } 0 \le \phi \le 1$$

Q: Do we need Lagrange Mult? No!

a) Take the derivative

$$\ell_{MAP}(\phi) = N_\alpha \log\phi + N_\beta \log(1-\phi)$$

where $N_\alpha \triangleq N_1 + \alpha - 1$

$N_\beta \triangleq N_0 + \beta - 1$

$$\frac{d\,\ell_{MAP}(\phi)}{d\phi} = \frac{N_\alpha}{\phi} + \frac{N_\beta}{\phi - 1}$$

b) Set to zero and solve

$$\frac{d\,\ell_{MAP}(\phi)}{d\phi} = 0 \Rightarrow N_\alpha(\phi-1) + N_\beta(\phi) = 0$$

$$\Rightarrow (N_\alpha + N_\beta)\phi = N_\alpha$$

$$\Rightarrow \phi = \frac{N_\alpha}{N_\alpha + N_\beta}$$

$$= \frac{N_1 + \alpha - 1}{N_1 + \alpha - 1 + N_0 + \beta - 1}$$

Intuition: proportional to # observed and hallucinated heads

Probabilistic View of Classification (Generative)

① ∃ some unknown dist. p^* parameterized by θ^* that generates labeled iid examples

$$(\vec{x}^{(i)}, y^{(i)}) \sim p^*(\vec{x}, y | \theta^*), \forall i$$

$$y^{(i)} \sim p(y | \theta^*)$$

$$\vec{x}^{(i)} \sim p(\vec{x} | y, \theta^*)$$

② Learning Algo.

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}} \log \prod_{i=1}^{N} p(\vec{x}^{(i)}, y^{(i)} | \theta) \; p(\theta)$$

$p(\theta)$: optional

③ Convert to a decision rule: output most likely label $\hat{y}$ for input $\vec{x}$

$$\hat{y} = h(\vec{x}) \triangleq \underset{y \in \{0,1\}}{\operatorname{argmax}} \; p(y | \vec{x}, \hat{\theta})$$

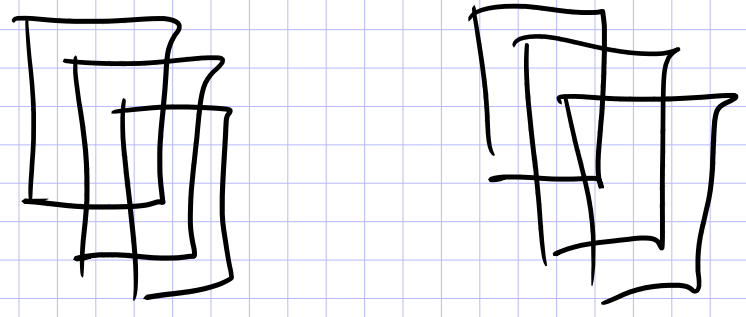
$$= \underset{y \in \{0,1\}}{\operatorname{argmax}} \; p(\vec{x} | y, \hat{\theta}) \, p(y | \hat{\theta})$$

Bayes' Rule

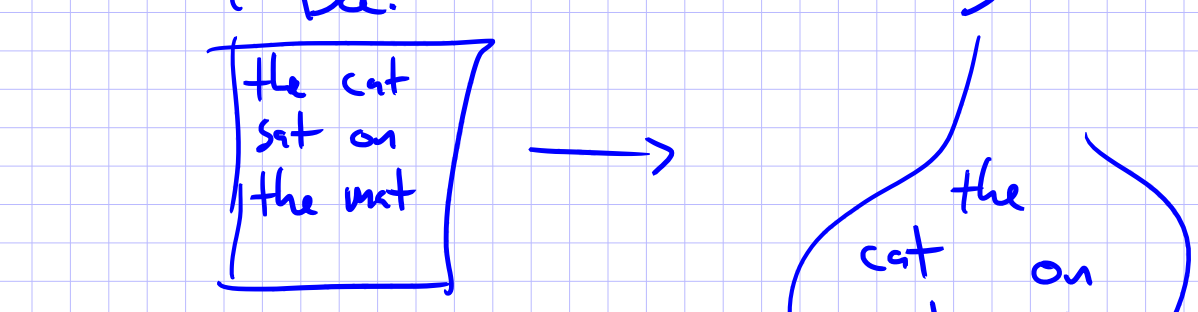
Natural Data

y (label) Econ. = 1 Onion = 0

$\vec{x}$ (words)

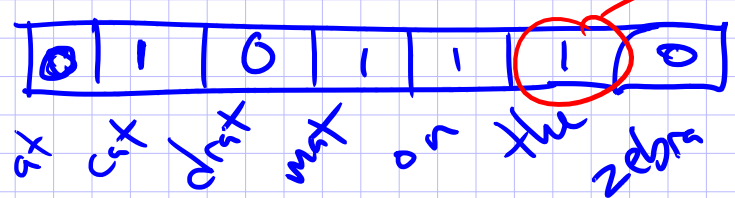


Conversion #1



Conversion #2

$\vec{x}^{(i)}$ =



Generate Synthetic Docs

$y \in \{0,1\}$ where Econ = 1 Onion = 0

$\vec{x} \in \{0,1\}^M$ where M = # of words in vocabulary

$\vec{x}^T = [x_1, \ldots, x_M]$

Generative Story:

$\phi \in \mathbb{R}$

$\theta = \begin{bmatrix} \theta_{10} & \theta_{11} \\ \theta_{20} & \theta_{21} \\ \vdots & \vdots \\ \theta_{M0} & \theta_{M1} \end{bmatrix}$

$y \sim \text{Bernoulli}(\phi) = p(y|\phi)$

$x_1 \sim \text{Bernoulli}(\theta_{1,y}) = p(x_1|y,\theta)$

$x_2 \sim \text{Bern.}(\theta_{2,y}) = p(x_2|y,\theta)$

$\vdots$

$x_M \sim \text{Bern}(\theta_{M,y}) = p(x_M|y,\theta)$

$$p(x_1, x_2, \ldots, x_M, y|\phi,\theta) = p(y|\phi)p(x_1|y,\theta)\cdots p(x_M|y,\theta)$$

Because each coin flip (for x_m) is conditionally independent given y

Naive Bayes Model

$$p(\vec{x}, y \mid \phi, \theta) = p(y \mid \phi) \prod_{m=1}^{M} p(x_m \mid y, \theta)$$