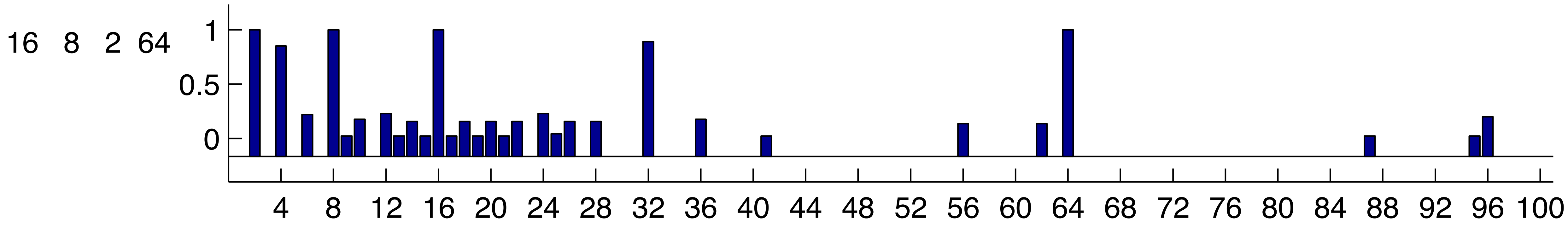
Chapter 3: Generative models for discrete data

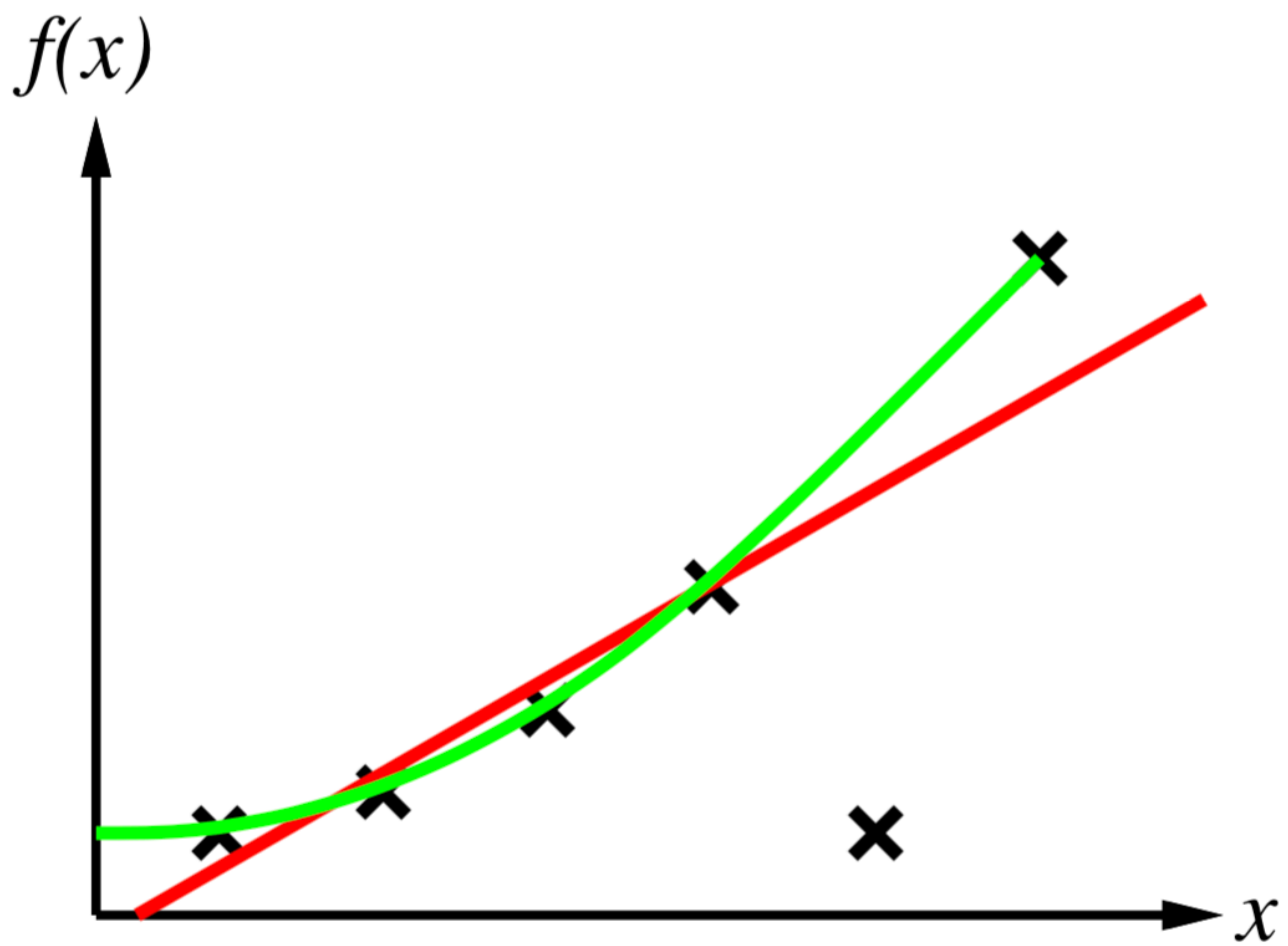
Mehdi	Lebdi	Southwest
Yifan	Li	East
Brandon	Lockhart	Back
Jialin	Lu	Back
Navaneeth	M.	Southwest
Arjun	Mahadevan	Northeast
Seyed Mohammad	Nourbakhsh	Northeast
Shuman	Peng	Back
SeyedHamed (Hamed)	RahmaniKhezri	Southwest
Rhea	Rodrigues	Southwest
Mohammadsadegh	Saberian	Northeast
Amir Hosein	Safari	Southeast
Bahar	Salamatian	West
Xiaoyu (Atticus)	Shi	West
Hamed	Shirzad	Middle
Neda	Shokraneh Kenary	Back
Xiangyu (Shawn)	Sun	Northwest
Chhavi	Verma	East
Lai	Wei	East
Andrew	Wesson	Northwest
Yi	Xie	Back
Ke (Jack)	Zhou	Southeast
Randall	Pyke	Middle

First	Last	W2
Niloufar	Abharigolsefidi	East
Mohammad Amin	Arab	Middle
Vahid Reza	Asadi	Southwest
Puria	Azadi Moghadam	Northwest
Adam	Banks	East
Evgeni (Eugene)	Borissov	Southeast
Logan	Born	West
Philip	Cho	Back
Peiyu	Cui	Middle
Adriano (Adrian)	D'Alessandro	Southwest
Ruizhi	Deng	Northwest
Mihir	Gajjar	Southeast
Atia	Hamidi Zadeh	Northeast
Fatemeh	Hasiri	West
Sha	Hu	Northwest
Xiang	Huang	Southeast
Salman	Imtiaz	West
Mohammadmahdi	Jahanara	Middle
Matthew	Jung	Northeast
Amogh	Kallihal	Middle
Arash	Khoeini	Southeast


OVERFITTING

Concept learning





Occam's razor: Prefer the simplest hypothesis

Posterior

$$p(h|\mathcal{D}) = \frac{p(\mathcal{D}|h)p(h)}{\sum_{h' \in \mathcal{H}} p(\mathcal{D}, h')}$$

Choosing a hypothesis

Maximum a posteriori (MAP) estimate:

$$
\begin{aligned}
h^{\mathrm{MAP}} &= \mathrm{argmax}_{h \in \mathcal{H}}\ p(h|\mathcal{D}) \\
&= \mathrm{argmax}_{h \in \mathcal{H}} \log p(\mathcal{D}|h) + \log p(h)
\end{aligned}
$$

Maximum likelihood estimate (MLE):

$$
h^{\mathrm{MLE}} = \mathrm{argmax}_{h \in \mathcal{H}} \log p(\mathcal{D}|h)
$$

Posterior predictive distribution

$$p(\tilde{y} = 1 | \tilde{x}, \mathcal{D}) = \sum_{h \in \mathcal{H}} p(\tilde{y} = 1 | \tilde{x}, h) p(h | \mathcal{D})$$

Exercise 2.6 Conditional independence

(Source: Koller.)

a. Let $H \in \{1, \ldots, K\}$ be a discrete random variable, and let e_1 and e_2 be the observed values of two other random variables E_1 and E_2. Suppose we wish to calculate the vector

$$\vec{P}(H|e_1, e_2) = (P(H = 1|e_1, e_2), \ldots, P(H = K|e_1, e_2))$$

Which of the following sets of numbers are sufficient for the calculation?

i. $P(e_1, e_2)$, $P(H)$, $P(e_1|H)$, $P(e_2|H)$

ii. $P(e_1, e_2)$, $P(H)$, $P(e_1, e_2|H)$

iii. $P(e_1|H)$, $P(e_2|H)$, $P(H)$

b. Now suppose we now assume $E_1 \perp E_2|H$ (i.e., E_1 and E_2 are conditionally independent given H). Which of the above 3 sets are sufficient now?

Show your calculations as well as giving the final result. Hint: use Bayes rule.

Steve Maine
@smaine

TIL that changing random stuff until your program works is "hacky" and "bad coding practice" but if you do it fast enough it's "#MachineLearning" and pays 4x your current salary

6:40 PM · 10 May 18

629 Retweets **1,692** Likes

Beta-binomial model

$$\mathrm{Bin}(k|n,\theta) \triangleq \binom{n}{k} \theta^k (1-\theta)^{n-k}$$

Binomial vs Bernoulli distributions

$$X_i \sim \mathrm{Ber}(\theta)$$

$$P(X_i|\theta) = \theta^{X_i}(1-\theta)^{1-X_i}$$

Beta-binomial model

$$\text{Beta}(\theta|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$$

Beta-binomial model

$$p(\theta|\mathcal{D}) \quad \propto \quad \mathrm{Bin}(N_1|\theta, N_0 + N_1)\mathrm{Beta}(\theta|a, b) \propto \mathrm{Beta}(\theta|N_1 + a, N_0 + b)$$

$$p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta) = \theta^{N_1}(1-\theta)^{N_0}\theta^{a}(1-\theta)^{b} = \theta^{N_1+a}(1-\theta)^{N_0+b}$$

Beta-binomial model

$$\hat{\theta}_{MLE} = \frac{N_1}{N}$$

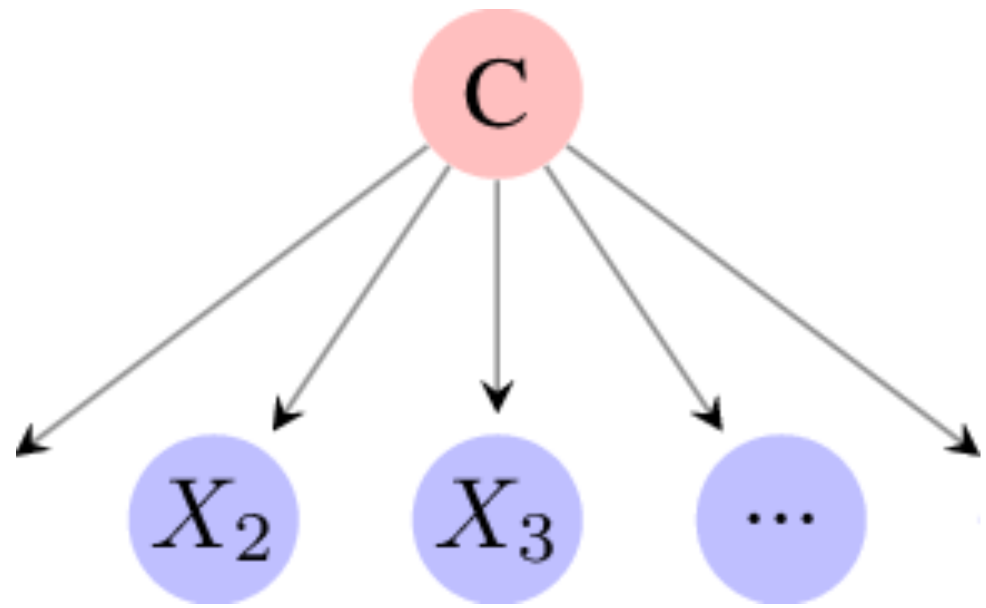
$$\overline{\theta} = \frac{a + N_1}{a + b + N}$$

Dirichlet-multinomial model

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{k=1}^{K} \theta_k^{N_k}$$

$$\mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \mathbb{I}(\mathbf{x} \in S_K)$$

Naive Bayes



$$p(\mathbf{x}|y = c, \boldsymbol{\theta}) = \prod_{j=1}^{D} p(x_j|y = c, \boldsymbol{\theta}_{jc})$$

Question 2.17

Suppose X, Y are two points sampled independently and uniformly at random from the interval [0,1]. What is the expected value of the lower value of the two?

Exercise 2.7 Pairwise independence does not imply mutual independence

We say that two random variables are pairwise independent if

$$p(X_2|X_1) = p(X_2) \tag{2.125}$$

and hence

$$p(X_2, X_1) = p(X_1)p(X_2|X_1) = p(X_1)p(X_2) \tag{2.126}$$

We say that n random variables are mutually independent if

$$p(X_i|X_S) = p(X_i) \quad \forall S \subseteq \{1, \ldots, n\} \setminus \{i\} \tag{2.127}$$

and hence

$$p(X_{1:n}) = \prod_{i=1}^{n} p(X_i) \tag{2.128}$$

Show that pairwise independence between all pairs of variables does not necessarily imply mutual independence. It suffices to give a counter example.