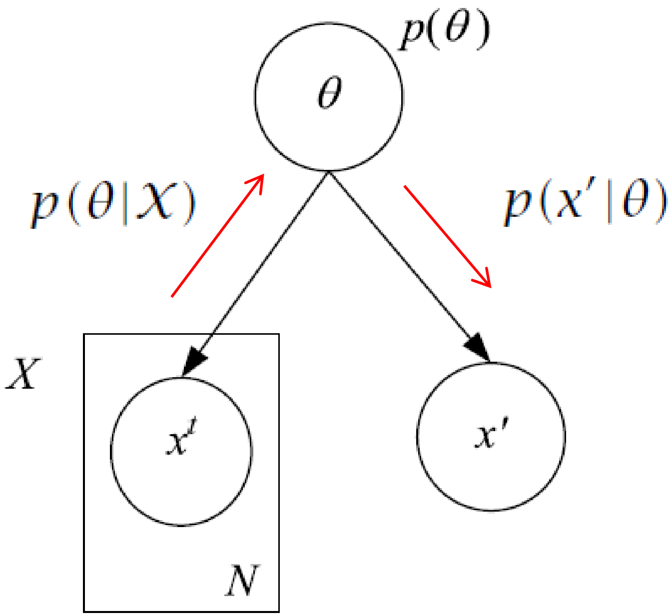
CHAPTER 16:

BAYESİAN ESTİMATİON
SECTION 4.4, 16.1, 16.2

Rationale

- Parameters θ not constant, but random variables with a prior, $p(\theta)$
- Bayes' Rule: $$p(\theta \mid X) = \frac{p(\theta)p(X \mid \theta)}{p(X)}$$

Generative Model



$$p(x', X, \theta) = p(\theta)p(X|\theta)p(x'|\theta)$$

$$\begin{aligned} p(x'|X) &= \frac{p(x',X)}{p(X)} = \frac{\int p(x',X,\theta)d\theta}{p(X)} = \frac{\int p(\theta)p(X|\theta)p(x'|\theta)d\theta}{p(X)} \\ &= \int p(x'|\theta)p(\theta|X)d\theta \end{aligned}$$

Bayesian Approach

$$p(x'|\mathcal{X}) = \int p(x'|\theta)p(\theta|\mathcal{X})d\theta$$

1. Prior $p(\theta)$ allows us to concentrate on region where θ is likely to lie, ignoring regions where it's unlikely
2. Instead of a single estimate with a single θ, we generate several estimates using several θ and average, weighted by how their probabilities

Even if prior $p(\theta)$ is uninformative, (2) still helps.

MAP estimator does not make use of (2):

$$\theta_{MAP} = \arg\max_{\theta} p(\theta|\mathcal{X})$$

Bayesian Approach

$$p(x'|X) = \int p(x'|\theta)p(\theta|X)d\theta$$

- In certain cases, it is easy to integrate
- Conjugate prior: Posterior $p(\theta|X)$ has the same parametric form as prior $p(\theta)$
- Sampling (Markov Chain Monte Carlo): Sample from the posterior and average
- Approximation: Approximate the posterior with a model easier to integrate
  - Laplace approximation: Use a Gaussian

Estimating the Parameters of a Distribution: Discrete case

- $r_i^t = 1$ if the *t-th* example has label i. Let the probability of label i be q_i.
- Sample likelihood (multinoulli distribution)

$$p(X \mid \mathbf{q}) = \prod_{t=1}^{N} \prod_{i=1}^{K} q_i^{r_i^t}$$

$$X = \{r_i^t : t = 1, \dots, N, \\ i = 1, \dots, K\}$$

- Dirichlet prior, α_i are hyperparameters

$$Dirichlet(\mathbf{q} \mid \boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_K)} \prod_{i=1}^{K} q_i^{\alpha_i - 1} = \frac{1}{B(\alpha)} \prod_{i=1}^{K} q_i^{\alpha_i - 1}$$

- Posterior

$$p(\mathbf{q} | X, \alpha) = \frac{\Gamma(\alpha_0 + N)}{\Gamma(\alpha_1 + N_1) \cdots \Gamma(\alpha_K + N_K)} \prod_{i=1}^{K} q_i^{\alpha_i + N_i - 1}$$

$$= Dirichlet(\mathbf{q} \mid \boldsymbol{\alpha} + \mathbf{n})$$

$$N_i : \#\{t : r_i^t = 1\}$$

$$\mathbf{n} = \begin{pmatrix} N_1 \\ \vdots \\ N_k \end{pmatrix}$$

- Dirichlet is a conjugate prior of multinoulli
  - prior α_i: pseudo-count. Its effect vanishes when Ni gets large.
  - Smoothing idea used in Lab 2
- With K=2, Dirichlet reduced to Beta distribution.

Dirichlet distribution

Probability

$$p(\mathbf{q}) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} q_i^{\alpha_i - 1}$$

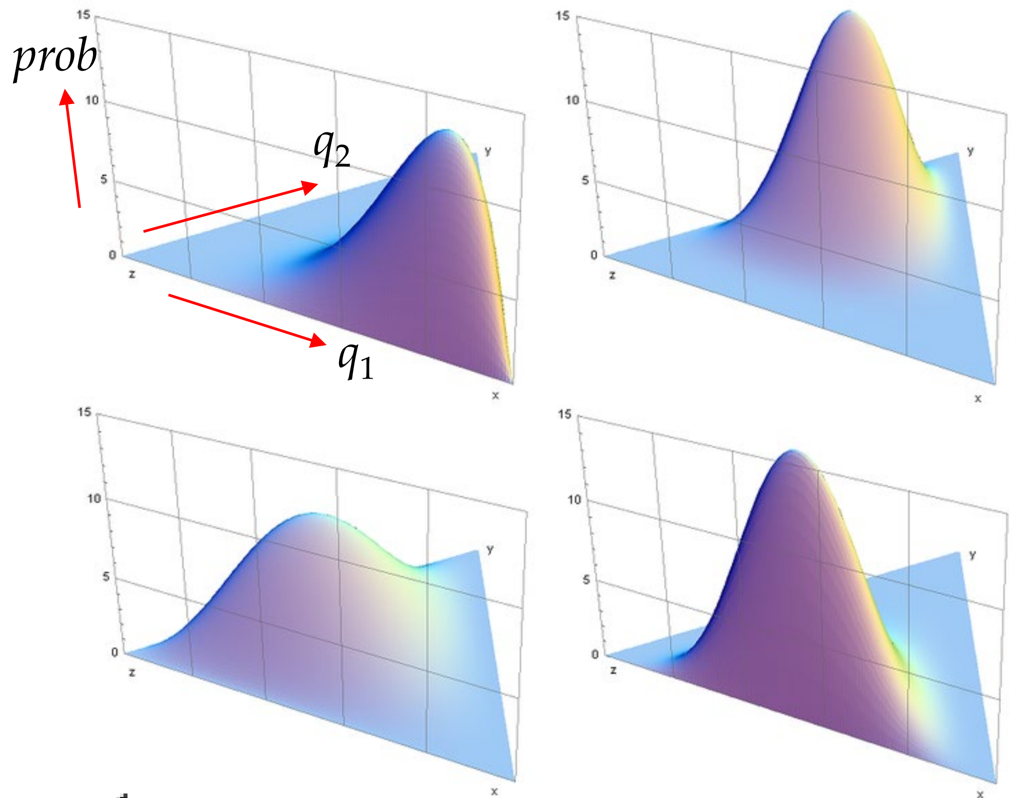
$$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$$

Mean

$$\mathbb{E}[q_i] = \frac{\alpha_i}{\sum_k \alpha_k}$$

Mode

$$q_i = \frac{\alpha_i - 1}{\sum_{i=1}^{K} \alpha_i - K}, \qquad \alpha_i > 1.$$



three classes

Bayesian Estimation for NB Predictions
1. For class prob $P(Y=C_i)$ (say, C classes)

First, multinoulli with Dirichlet prior:

$Y \sim \text{Multinoulli}(\mathbf{q})$ (i.e. $P(Y=C_i) = q_i$)

$\mathbf{q} \sim \text{Dirichlet}(\boldsymbol{\alpha})$

$X = \{N_1, N_2, \ldots N_C\}$

$P(\mathbf{q} \mid X) \sim \text{Dirichlet}(\alpha_1+N_1, \alpha_2+N_2, \ldots, \alpha_C+N_C)$

$$P(Y = C_i \mid X) = \int P(Y = C_i \mid \mathbf{q})\, P(\mathbf{q} \mid X)\, d\mathbf{q}$$

$$= \int q_i\, P(\mathbf{q} \mid X)\, dq_j$$

$$= E[q_i \mid X] = (a_i + N_i) / (a_0 + N)$$

Bayesian Estimation for NB Predictions

2. Conditional Distribution P($X_j = v_k \mid C_i$)

Same rationale!

Recall $$p_{ijk} \equiv p(z_{jk} = 1 \mid C_i) = p(x_j = v_k \mid C_i)$$

In Lab 2: x_j is the j-th word in a message, v_k is the k-th word of a dictionary

Assume: 1. Each x_j can take value in $\{v_1, \ldots, v_K\}$ (K possible values/dict words)

2. Drop j from p_{ijk} if all $\{x_j \mid C_i : j\}$ share the "same" distribution

Likelihood (for one example/document):

$$p(\mathbf{x}|C_i) = \prod_{j=1}^{d} \prod_{k=1}^{K} p_{ik}^{z_{jk}}$$

Assume Dirichlet prior

$$\theta_i = \begin{pmatrix} \theta_{i,1} \\ \vdots \\ \theta_{i,K} \end{pmatrix}$$

$$\mathbf{p}_i := (p_{i1}, \ldots, p_{iK}) \sim \text{Dirichlet}(\theta_i)$$

Then the posterior is

$$\mathbf{p}_i|X \sim \text{Dirichlet}(\theta_i + n_i) \quad \text{where} \quad n_i = \begin{pmatrix} \#\{x_j^t = 1 : \text{all } t, j, r_i^t = 1\} \\ \vdots \\ \#\{x_j^t = K : \text{all } t, j, r_i^t = 1\} \end{pmatrix}$$