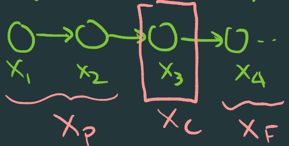
Last 2 classes

- Beta-Bernoulli model (for learning 'p' from Ber(p) data)
- Decision theory – MAP estimator, posterior prediction, credible intervals
- Bayes Nets – Graphical 'language' for complex probabilistic models (DAG which represents independence rel^{n}s in your model)
- Conditional independence $\equiv$ d-separation

(are $X_A \perp\!\!\!\perp X_B \mid X_C$
$X_A \equiv \{X_i \mid i \in A\}$)

Today

- Dirichlet priors
- Naive Bayes classifier
- Gaussian models



Markov – $X_F \perp\!\!\!\perp X_P \mid X_C$

(Bishop, Chapter 8, sec 1,2 (?))

multiclass data (clustering / classification)

count of word i in data

- data $D = \{X_1, X_2, \ldots, X_n\} \in \{1, 2, \ldots, K\}^n$ ← number of words
  eg - set of english words
- for $i \in [K]$, data D contains N_i copies of type i
- model $\mathcal{M}$: X_i generated i.i.d. from $Multinomial(\theta_1, \theta_2, \ldots, \theta_K)$ distn

Eg - email filtering, classes = {spam, not spam}

sentiment analysis

datasets ≡ 'bag of words' rep^{n} of an email

← $\underbrace{\text{Each word is of type } k \text{ w.p } \frac{\theta_k}{\sum \theta_i} \text{ iid}}_{\text{NAIVE}}$

$Mult(n, \theta_1, \theta_2, \ldots, \theta_K)$

Eg - If $K = 2$, then $Bin(n, p)$

ie, $\mathcal{A} = \{0,1\}$ $= Mult(n, p, 1-p)$

$$\boxed{\mathcal{L}(N_1, N_2, \ldots, N_K \mid \theta) = \prod_{i=1}^{K} \theta_i^{N_i}}$$

Multinomial Likelihood

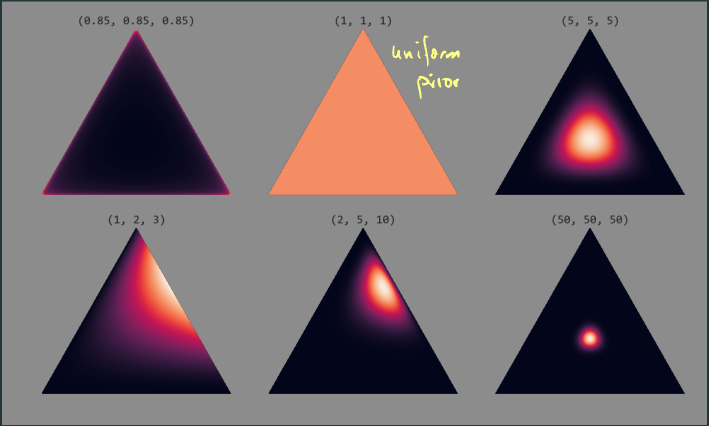
the Dirichlet distribution

Dirichlet distribution

K-simplex

- $x \in \{x_i \in [0,1], \sum_{i=1}^{K} x_i = 1\}$, parameters: $\Theta = (\alpha_1, \alpha_2, \ldots, \alpha_K)$
- pdf: $p(x) \propto \prod_{i=1}^{K} x_i^{\alpha_i - 1}$

hyperparameters $\alpha_i \geq 0$



the Dirichlet distribution

Dirichlet distribution

- $x \in \{x_i \in [0,1], \sum_{i=1}^{K} x_i = 1\}$, parameters: $\Theta = (\alpha_1, \alpha_2, \ldots, \alpha_K)$
- denote $\alpha = \{\alpha_i\}_{i=1}^{K})$; Dirichlet pdf

$\Gamma(\alpha_1+\alpha_2+\ldots+\alpha_K) / \Gamma(\alpha_1)\Gamma(\alpha_2)\ldots\Gamma(\alpha_K)$

$$p(x) = \frac{1}{Z(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i - 1}$$

- normalizing constant: $\frac{1}{Z(\alpha)} = \frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\prod_{i=1}^{K} \Gamma(\alpha_i)}$
  (Dirichlet norm'd const)

- Recall. $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$, $\Gamma(1) = 1$

  $\Gamma(k) = (k-1)!$ if $k \in \mathbb{N}$

multiclass data and Dirichlet priors

- for $i \in [K]$, data D contains N_i copies of type i
- model $\mathcal{M}$: X_i generated i.i.d. from $Multinomial(\theta_1, \theta_2, \ldots, \theta_K)$ distn

Dirichlet-Multinomial model

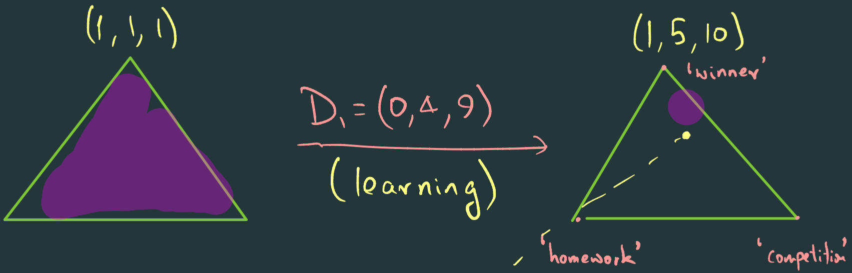
- prior parameters: $\Theta_0 = (\alpha_1, \alpha_2, \ldots, \alpha_K) \in \mathbb{R}_+^K$ (hyperparameters)
- Dirichlet prior: $Dir(\alpha_1, \alpha_2, \ldots, \alpha_K) \sim p(\theta) \propto \prod_{i=1}^K \theta_i^{\alpha_i - 1}$
- likelihood: $p(D|\theta) = \prod_{i=1}^K \theta_i^{N_i}$
- posterior: $p(\theta|D) \sim Dir(\alpha_1 + N_1, \alpha_2 + N_2, \ldots, \alpha_K + N_K)$
- marginal likelihood: let $m = \sum_{i=1}^K \alpha_i$

posterior = $\dfrac{\text{prior} \times \text{likelihood}}{\text{marginal likelihood}}$

$$p(D) = \frac{\Gamma(m)}{\Gamma(n+m)} \prod_{i=1}^{K} \frac{\Gamma(N_i + \alpha_i)}{\Gamma(\alpha_i)}$$

Spam filtering

- Consider only spam emails : $D_1, D_2, \dots, D_{n_s}$
- $D_i \equiv (N_1^i, N_2^i, \dots, N_K^i) \equiv$ word counts



- Now if we get a new email D', what is $\mathbb{P}[D' \mid \text{spam}]$

$$\propto \prod_{i=1}^{K} (\alpha_i^s)^{N_i'}$$

$$= \frac{1}{Z(1,5,10)} (1)^{N_1'} (5)^{N_2'} (10)^{N_3'}$$

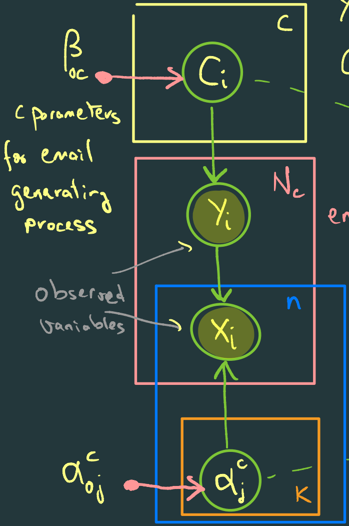
Naive Bayes classifier

- For spam – take data $D^s_1, D^s_2, \dots, D^s_{n_s}$
  – learn posterior $\mathrm{Dir}(\alpha^s_1, \alpha^s_2, \dots, \alpha^s_K)$
- For non-spam (regular) – take data $D^r_1, D^r_2, \dots, D^r_{n_r}$
  – learn posterior $\mathrm{Dir}(\alpha^r_1, \alpha^r_2, \dots, \alpha^r_K)$

- Given new email $D' := (N_1, N_2, \dots, N_K)$ is it regular or spam?
  – Compute $p(D' \mid \text{spam}) = \dfrac{\prod_{i=1}^{K} (\alpha^s_i)^{N_i}}{Z(\{\alpha^s_i\})}$, $p(D' \mid \text{regular}) = \dfrac{\prod_{i=1}^{K} (\alpha^r_i)^{N_i}}{Z(\{\alpha^r_i\})}$
  – $p(\text{spam} \mid D') \propto p(D' \mid \text{spam}) \underbrace{p(\text{spam})}_{?}$, $p(\text{regular} \mid D') \propto p(D' \mid \text{reg}) \underbrace{p(\text{reg})}_{?}$

need model for class label

Naive Bayes model for spam classification



'learning'

$Y_i = \{\text{spam, regular}\}$

$C_i \equiv$ parameters of email generation process

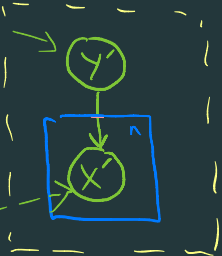
$X_i \equiv (N_1, N_2, \ldots, N_K)$

emails (datasets)

N_c examples for each class

n words in each email

K parameters for each class (one for each word)



'prediction'

generative models for continuous data