# Probabilistic Machine Learning: 5. Dirichlet-multinomial model

Tomasz Kajdanowicz, Przemysław Kazienko (substitution)

Department of Computational Intelligence
Wroclaw University of Technology



HR EXCELLENCE IN RESEARCH

Wrocław University of Science and Technology

The presentation was inspired by Chapter 3 of Kevin Murphy book "Machine Learning A Probabilistic Perspective", 2012, MIT.

Apropriate agreements to propagate his ideas has been acquired.

# Already: coin toss problem

We have covered:

- inferring a distribution over a discrete variable drawn from a finite hypothesis space
- inferring the probability that a coin shows up heads
- given a series of discrete observations

Let’s focus now on a dice:

- dice is $K$ sided :)

# Johann Dirichlet (1805–1859)

- German mathematician with French roots
- with the support of Humboldt and Gauss, Dirichlet was offered a teaching position at the University of Breslau (1827-1828)
- in 1842 obtained a full professor position at the University of Breslau
- Dirichlet distribution named after him

# Dirichlet distribution

- multivariate generalization of the beta distribution
- parameters:
  - $K > 2$ - categories
  - $\alpha_1, \ldots, \alpha_K$ - concentration parameters ($\alpha_i > 0$)
- support: $x_1, \ldots, x_K$ where $x_i \in (0, 1)$ and $\sum_{i=1}^{K} x_i = 1$

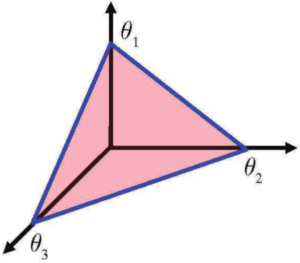
# Dirichlet distribution

About the support ($x$):

- over the probability simplex
- $S_K = \{x : 0 \leq x_k \leq 1, \sum_{k=1}^{K} x_k = 1\}$

Probability distribution over $X$

- $Dir(x|\alpha) = \frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \prod_{k=1}^{K} x_k^{\alpha_k - 1}$
- with normalizing factor:
  $\mathrm{B}(\boldsymbol{\alpha}) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}$



# Dirichlet distribution

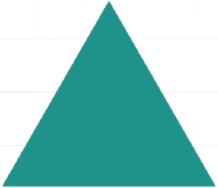


Figure: Dir(1,1,1)

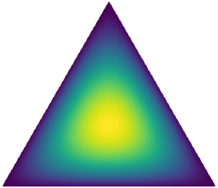


Figure: Dir(2,2,2)

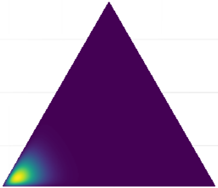


Figure: Dir(20,2,2)

# Dirichlet distribution

- mean:

$$\mathbb{E}(X_i) = \frac{\alpha_i}{\sum_k \alpha_k}$$

- mode:

$$x_i = \frac{\alpha_i - 1}{\sum_{k=1}^{K} \alpha_k - K}, \quad \alpha_i > 1$$

# Recipe of specifying the model

Define

- likelihood
- prior

and derive

- posterior
- posterior predictive

# Likelihood

Let's suppose:

- we observe $N$ dice rolls
- $\mathcal{D} = \{x_1, x_2, \ldots, x_N\}$
- $x_i \in \{1, \ldots, K\}$

Assuming the data is iid, likelihood has the form:

$$\mathcal{L}(\mathcal{D} \mid \theta) = \prod_{k=1}^{K} \theta_k^{N_k}$$

The number of times event $k$ occured is given by sufficient statistics: $N_k = \sum_{i=1}^{N} \mathbb{I}(y_i = k)$

# What about prior?

- need of prior from $K$-dimensional probability simplex
- the best should be conjugate one
- fortunatelly, Dirichlet distribution satisfies both criteria

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

**Conjugate priors**

When the prior and the posterior have the same form, we say that the prior is a conjugate prior for the corresponding likelihood.

# Posterior

**Multiply the likelihood by the prior:**

$$
\begin{aligned}
p(\theta|D) &\propto p(\mathcal{D} \mid \theta)p(\theta) \\
&\propto \prod_{k=1}^{K} \theta_k^{N_k} \theta_k^{\alpha_k - 1} = \prod_{k=1}^{K} \theta_k^{\alpha_k + N_k - 1} \\
&= Dir(\theta \mid \alpha_1 + N_1, \ldots, \alpha_k + N_K)
\end{aligned}
$$

- posterior is obtained by adding the prior hyper-parameters (pseudo-counts) $\alpha_k$ to the empirical counts $N_k$
- MAP estimate is given by $\hat{\theta}_k = \frac{N_k + \alpha_k - 1}{N + \sum_{k=1}^{K} \alpha_k - K}$

# Posterior predictive distribution

How to make prediction of future observable data?

Predicting the probability of single toss under posterior:

$$
\begin{aligned}
p(X = j|\mathcal{D}) &= \int p(X = j|\theta)p(\theta|\mathcal{D})d\theta \\
&= \int p(X = j|\theta_j)\left[\int p(\theta_{-j}, \theta_j|\mathcal{D})d\theta_{-j}\right]d\theta_j \\
&= \int \theta_j p(\theta_j|\mathcal{D})d\theta_j = \mathbb{E}(\theta_j|\mathcal{D}) \\
&= \frac{\alpha_j + N_j}{\sum_{k=1}^{K} \alpha_k + N_k}
\end{aligned}
$$

# Example: language model using bag of words

- application of Bayesian smoothing using the Dirichlet-multinomial model is to language modeling
- predict which words might occur next in a sequence

Let's assume:

- the *i*'th word, $X_i \in \{1, \ldots, K\}$, is sampled independently from all the other words using a $Cat(\theta)$ distribution
- Cat - categorical distribution (Multinouli, generalization of Bernouli)
  - when the *i*-th outcome is obtained, the *i*-th entry of the random variable *X* takes value 1, while all other entries take value 0, e.g. [0,1,0,0,0,0,0]
- called: **bag of word model**

# Example: language model using bag of words

## Example

Mary had a little lamb, little lamb, little lamb,
Mary had a little lamb, its fleece as white as snow

## Vocabulary

| mary | lamb | little | big | fleece | white | black | snow | rain | unk |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |