Introduction to topic models: Building up towards LDA

Vivi Nastase

Summer semester 2012

ICL, University of Heidelberg

Motivation

Assumption: the data was generated according to some model that we try to retrieve.

Motivation

Assumption: the data was generated according to some model that we try to retrieve.

The **observation** $\mathcal{O}$: the texts.
The **model** μ: parameters and probability distributions that generated the data – the one that fits the data best.

$$p(\mathcal{O}, \mu) = \underbrace{p(\mathcal{O}|\mu)}_{\textbf{likelihood}} \underbrace{p(\mu)}_{\textbf{prior}} = \underbrace{p(\mu|\mathcal{O})}_{\textbf{posterior}} \underbrace{p(\mathcal{O})}_{\textbf{evidence}}$$

Motivation

Assumption: the data was generated according to some model that we try to retrieve.

The **observation** $\mathcal{O}$: the texts.
The **model** μ: parameters and probability distributions that generated the data – the one that fits the data best.

$$p(\mathcal{O}, \mu) = \underbrace{p(\mathcal{O}|\mu)}_{\textbf{likelihood}} \underbrace{p(\mu)}_{\textbf{prior}} = \underbrace{p(\mu|\mathcal{O})}_{\textbf{posterior}} \underbrace{p(\mathcal{O})}_{\textbf{evidence}}$$

Best fit:

- maximize likelihood
- maximize the posterior
- compute the probability distribution of the posterior

Maximum likelihood estimation (ML)

Likelihood:

$$L(\mu|\mathcal{O}) = p(\mathcal{O}|\mu) = \prod_{x \in \mathcal{O}} p(x|\mu)$$

$$\mathcal{L}(\mu|\mathcal{O}) = \log L(\mu|\mathcal{O}) = \sum_{x \in \mathcal{O}} \log p(x|\mu)$$

Maximum likelihood estimation (ML)

Likelihood:

$$L(\mu|\mathcal{O}) = p(\mathcal{O}|\mu) = \prod_{x \in \mathcal{O}} p(x|\mu)$$

$$\mathcal{L}(\mu|\mathcal{O}) = \log L(\mu|\mathcal{O}) = \sum_{x \in \mathcal{O}} \log p(x|\mu)$$

The model that maximizes the likelihood:

$$\mu_{ML} = argmax_{\mu} \mathcal{L}(\mu|\mathcal{O}) = argmax_{\mu} \sum_{x \in \mathcal{O}} \log p(x|\mu)$$

Maximum likelihood estimation (ML)

Likelihood:

$$L(\mu|\mathcal{O}) = p(\mathcal{O}|\mu) = \prod_{x \in \mathcal{O}} p(x|\mu)$$

$$\mathcal{L}(\mu|\mathcal{O}) = \log L(\mu|\mathcal{O}) = \sum_{x \in \mathcal{O}} \log p(x|\mu)$$

The model that maximizes the likelihood:

$$\mu_{ML} = argmax_{\mu} \mathcal{L}(\mu|\mathcal{O}) = argmax_{\mu} \sum_{x \in \mathcal{O}} \log p(x|\mu)$$

Maximum:

$$\frac{\partial \mathcal{L}(\mu|\mathcal{O})}{\partial \mu_k} = 0; \forall \mu_k \in \mu$$

Maximum likelihood estimation (ML) – example

$\mathcal{O}$: N coin tossing experiments, $N = n_h + n_t$

Maximum likelihood estimation (ML) – example

$\mathcal{O}$: N coin tossing experiments, $N = n_h + n_t$

μ: p_h $\quad p(X = x_i|p_h) = p_h^{x_i}(1 - p_h)^{1-x_i}$

$x_i = 1$ for heads, $x_i = 0$ for tails

Maximum likelihood estimation (ML) – example

$\mathcal{O}$: N coin tossing experiments, $N = n_h + n_t$

μ: p_h $\quad p(X = x_i|p_h) = p_h^{x_i}(1 - p_h)^{1-x_i}$

$x_i = 1$ for heads, $x_i = 0$ for tails

$$\mathcal{L}(p_h|N) = \sum_{i=1,N} \log p(X = x_i|p_h)$$

Maximum likelihood estimation (ML) – example

$\mathcal{O}$: N coin tossing experiments, $N = n_h + n_t$

μ: p_h $\quad p(X = x_i|p_h) = p_h^{x_i}(1 - p_h)^{1-x_i}$

$x_i = 1$ for heads, $x_i = 0$ for tails

$$
\begin{aligned}
\mathcal{L}(p_h|N) &= \sum_{i=1,N} \log p(X = x_i|p_h) \\
&= \sum_{i=1,N} \log(p_h^{x_i}(1 - p_h)^{1-x_i}) \\
&= \sum_{i=1,N} x_i \log p_h + (1 - x_i) \log(1 - p_h) \\
&= n_h \log p_h + n_t \log(1 - p_h)
\end{aligned}
$$

$$
\begin{aligned}
p_{h\ ML} &= argmax_p \mathcal{L}(p_h|N) \\
&\rightarrow \frac{\partial \mathcal{L}(p_h|N)}{\partial p_h} = \frac{n_h}{p_{h\ ML}} - \frac{n_t}{1 - p_{h\ ML}} = 0
\end{aligned}
$$

Maximum a posteriori estimation (MAP)

Similar to ML estimation, but incorporates some prior knowledge about the parameters.

$$p(\mathcal{O}, \mu) = \underbrace{p(\mathcal{O}|\mu)}_{\textbf{likelihood}} \underbrace{p(\mu)}_{\textbf{prior}} = \underbrace{p(\mu|\mathcal{O})}_{\textbf{posterior}} \underbrace{p(\mathcal{O})}_{\textbf{evidence}}$$

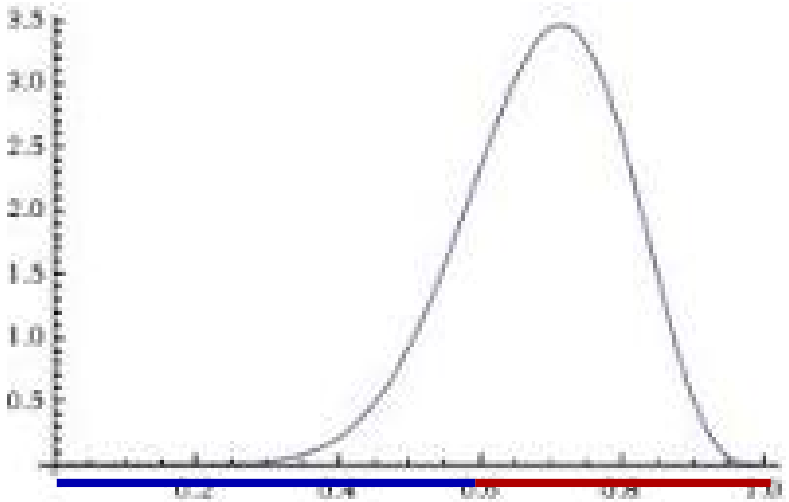
$$p(\mu|\mathcal{O}) = \frac{p(\mathcal{O}|\mu)\ p(\mu)}{p(\mathcal{O})}$$

Maximum a posteriori estimation (MAP)

Similar to ML estimation, but incorporates some prior knowledge about the parameters.

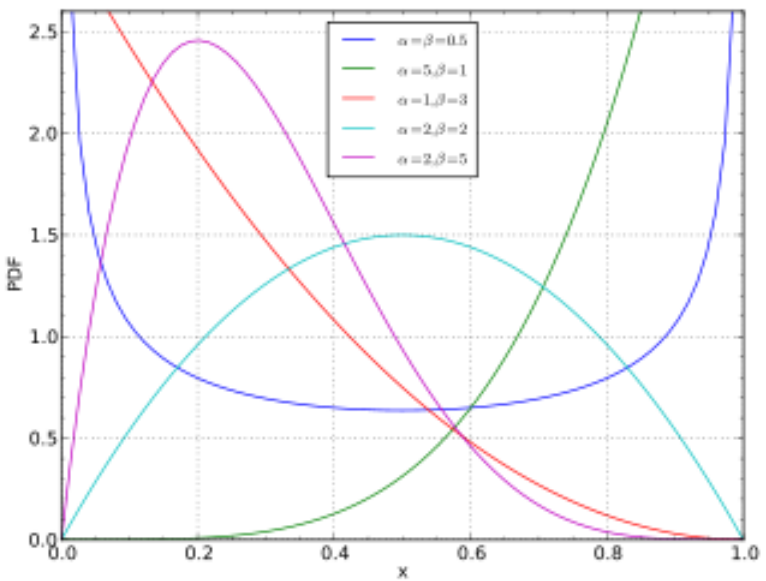
$$
\begin{aligned}
\mu_{MAP} &= \mathit{argmax}_\mu p(\mu|\mathcal{O}) \\
&= \mathit{argmax}_\mu \frac{p(\mathcal{O}|\mu)p(\mu)}{p(\mathcal{O})} \\
&= \mathit{argmax}_\mu \underbrace{p(\mathcal{O}|\mu)}_{\text{likelihood}} \underbrace{p(\mu)}_{\text{prior}} \\
&= \mathit{argmax}_\mu \mathcal{L}(\mathcal{O}|\mu) + \log p(\mu)
\end{aligned}
$$

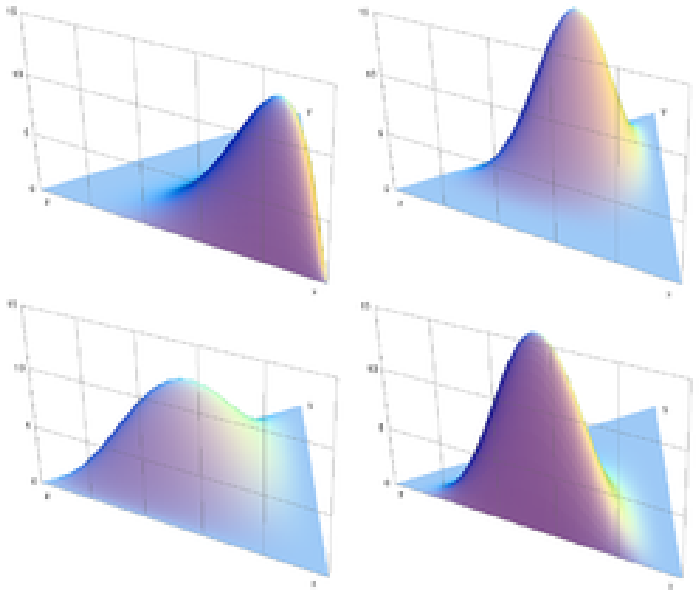
About priors – Beta/Dirichlet



$$p(p_h|\alpha_h, \alpha_t) = Beta(p_h|\alpha_h, \alpha_t) = \frac{1}{B(\alpha_h, \alpha_t)} p_h^{\alpha_h - 1}(1 - p_h)^{\alpha_t - 1}$$

About priors – Beta/Dirichlet





$$p(p_h|\alpha_h,\alpha_t) = Beta(p_h|\alpha_h,\alpha_t) = \frac{1}{B(\alpha_h,\alpha_t)} p_h^{\alpha_h-1}(1-p_h)^{\alpha_t-1}$$

Maximum a posteriori estimation (MAP) – example

N coin tossing experiments, head/tails; $N = n_h + n_t$
p_h – probability to get a head
Prior belief about the coin: $p(p_h) = Beta(p_h|\alpha_h, \alpha_t)$

$$\begin{aligned}
\mu_{MAP} &= argmax_\mu \mathcal{L}(p_h|N) + \log p(p_h) \\
&\rightarrow \frac{n_h}{p_h} - \frac{n_t}{1-p_h} + \frac{\alpha_h - 1}{p_h} - \frac{\alpha_t - 1}{1-p_h} = 0 \\
\mu_{MAP} &= \frac{n_h + \alpha_h - 1}{N + \alpha_h - 1 + \alpha_t - 1}
\end{aligned}$$

Bayesian estimation