1 Summary

This is a note for [2]. Let's start from Naive Bayes classifier.

$$P(word = w|class = k) = \frac{\exp(\eta_{kw})}{\sum_v \exp(\eta_{kv})}$$

The important part is $\eta_{kw} \in \mathbb{R}$, which allows η_{kw} to take negative values too. If we directly model the probability $p_{kw} = \frac{\#\text{ word } w \text{ in class } k}{\#\text{ words in class } k}$, then it has the limitation that 1) $1 \geq p_{kw} \geq 0$ and 2) $\int_\Omega p_{kw} = 1$. This technique is called "natural parameterization", which is denoted as $\eta_{kw} = \log(p_{kw}) + c$ where c is a constant value. Also note that log-linear models are easy when because the log transformation makes the objective function as a sum and it is easy to take the derivatives when compared to a product.

$$P(w|y_d, m, \eta) = \frac{\exp(m + \eta_{y_d})}{\sum_i (m_i + n_{y_d,i})}$$

Now, the prior is no longer limited to Dirichlet distribution. For example, a prior can be a Gaussian distribution $\boldsymbol{\eta} \sim N(0, \Sigma)$ where Σ can encode the similarity between words.

In the SAGE paper [2], Laplace prior or double exponential distribution (also see the graph at [1]) is used to produce sparsity. Sparsity also encourages interpretability because we can now focus on fewer parameters (or words).

References

[1] Why is Laplace prior producing sparse solutions? `http://stats.stackexchange.com/questions/177210/why-is-laplace-prior-producing-sparse-solutions/`. [Online; accessed 9-Sep-2016].

[2] Jacob Eisenstein, Amr Ahmed, and Eric P. Xing. Sparse additive generative models of text. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1041–1048, 2011.