Statistical Machine Learning
Homework2

方言

2021210929

2022 年 4 月 19 日

1 Ensemble

1.1 Problem 1

1.1.1 Prove the optimal solution using Lagrange multiplier.

证明:

目标函数:

$$\begin{aligned}L(\boldsymbol{\theta}) &= \log p(\boldsymbol{t}|\boldsymbol{\theta}) \\ &= \sum_{n=1}^{N} \log(\sum_{k=1}^{K} \pi_k y_{nk}^{t_n}[1-y_{nk}]^{1-t_n})\end{aligned}$$

给定约束条件 $\sum_{k=1}^{K}\pi_k = 1$，构造拉格朗日函数:

$$\begin{aligned}F(\boldsymbol{\pi},\lambda) &= L(\boldsymbol{\theta}) + \lambda(\sum_{k=1}^{K}\pi_k - 1) \\ &= \sum_{n=1}^{N} \log(\sum_{k=1}^{K} \pi_k y_{nk}^{t_n}[1-y_{nk}]^{1-t_n}) + \lambda(\sum_{k=1}^{K}\pi_k - 1)\end{aligned}$$

将其对 π_k 求偏导，并且令偏导为 0，可得:

$$\frac{\partial F(\boldsymbol{\pi},\lambda)}{\partial \pi_k} = \sum_{n=1}^{N} \frac{y_{nk}^{t_n}[1-y_{nk}]^{1-t_n}}{\sum_j \pi_j y_{nj}^{t_n}[1-y_{nj}]^{1-t_n}} + \lambda = 0,\ 1 \le k \le K$$

将其对 λ 求偏导，并且令偏导为 0，可得:

$$\frac{\partial F(\boldsymbol{\pi},\lambda)}{\partial \lambda} = \sum_{k=1}^{K}\pi_k - 1 = 0$$

令 $\gamma_{nk} = \frac{\pi_k y_{nk}^{t_n}[1-y_{nk}]^{1-t_n}}{\sum_j \pi_j y_{nj}^{t_n}[1-y_{nj}]^{1-t_n}}$，可得:

$$\begin{aligned}\sum_{k=1}^{K}\pi_k \frac{\partial F(\boldsymbol{\pi},\lambda)}{\partial \pi_k} &= \sum_{k=1}^{K}\sum_{n=1}^{N}\gamma_{nk} + \sum_{k=1}^{K}\pi_k\lambda \\ &= \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma_{nk} + \lambda\sum_{k=1}^{K}\pi_k \\ &= N + \lambda = 0\end{aligned}$$

代入 $\lambda = -N$，可得:

$$\sum_{n=1}^{N} \frac{\gamma_{nk}}{\pi_k} - N = 0$$

即:

$$\pi_k = \frac{\sum_{n=1}^{N} \gamma_{nk}}{N},\ 1 \le k \le K$$

由此，即可证明 $\{\pi_k\}_{k=1}^{K}$ 满足上式

1.1.2 Fix $\{\pi_k\}_{k=1}^{K}$, prove that

$$\nabla_{\omega_k} L = \sum_{n=1}^{N} \gamma_{nk}(t_n - y_{nk})\phi_n$$

证明:

由于:

$$\sigma(x) = \frac{1}{1+e^{-x}}$$
$$\sigma'(x) = \frac{e^x}{(e^x+1)^2} = \sigma(x)(1-\sigma(x))$$

将目标函数 L 对 ω_k 求偏导，可得:

$$\begin{aligned}
\nabla_{\omega_k} L &= \sum_{n=1}^{N} \pi_k \frac{t_n y_{nk}^{t_n-1}(1-y_{nk})^{1-t_n} + (t_n-1)y_{nk}^{t_n}(1-y_{nk})^{-t_n}}{\sum_j \pi_j y_{nj}^{t_n}[1-y_{nj}]^{1-t_n}} \cdot (1-y_{nk})y_{nk} \cdot \phi_n \\
&= \sum_{n=1}^{N} [t_n(1-y_{nk}) + (t_n-1)y_{nk}] \cdot \frac{\pi_k y_{nk}^{t_n}[1-y_{nk}]^{1-t_n}}{\sum_j \pi_j y_{nj}^{t_n}[1-y_{nj}]^{1-t_n}} \cdot \phi_n \\
&= \sum_{n=1}^{N} \gamma_{nk}(t_n - y_{nk})\phi_n
\end{aligned}$$

由此得证

1.1.3 Caculate $H_k = -\nabla_{\omega_k}\nabla_{\omega_k} L$

令 $z_{nk} = y_{nk}^{t_n}[1-y_{nk}]^{1-t_n}$，首先计算:

$$\begin{aligned}
\nabla_{\omega_k}\gamma_{nk} &= \nabla_{\omega_k}\left(\frac{\pi_k z_{nk}}{\sum_j \pi_j z_{nj}}\right) \\
&= \frac{\pi_k(\sum_j \pi_j z_{nj} - \pi_k z_{nk})}{(\sum_j \pi_j z_{nj})^2} \cdot \nabla_{\omega_k} z_{nk}
\end{aligned}$$

由于:

$$\begin{aligned}
\nabla_{\omega_k} z_{nk} &= \nabla_{\omega_k}\left(y_{nk}^{t_n}[1-y_{nk}]^{1-t_n}\right) \\
&= y_{nk}^{t_n-1}[1-y_{nk}]^{-t_n} \cdot (t_n - y_{nk}) \cdot \nabla_{\omega_k} y_{nk} \\
&= z_{nk}(t_n - y_{nk})\phi_n^T
\end{aligned}$$

由此可得:

$$
\begin{aligned}
H_k &= -\nabla_{\omega_k}\nabla_{\omega_k}L \\
&= \nabla_{\omega_k}\left(\sum_{n=1}^{N}\gamma_{nk}(y_{nk}-t_n)\phi_n\right) \\
&= \sum_{n=1}^{N}\left[(y_{nk}-t_n)\phi_n\cdot\nabla_{\omega_k}\gamma_{nk}+\gamma_{nk}\phi_n y_{nk}(1-y_{nk})\cdot\nabla_{\omega_k}y_{nk}\right] \\
&= \sum_{n=1}^{N}\left[(y_{nk}-t_n)\phi_n\cdot\frac{\pi_k z_{nk}(\sum_j\pi_j z_{nj}-\pi_k z_{nk})}{(\sum_j\pi_j z_{nj})^2}\cdot(t_n-y_{nk})\phi_n^T+\gamma_{nk}y_{nk}(1-y_{nk})\cdot\phi_n\phi_n^T\right] \\
&= \sum_{n=1}^{N}\left[(y_{nk}-t_n)\phi_n\cdot\frac{\pi_k z_{nk}}{\sum_j\pi_j z_{nj}}\cdot\frac{\sum_j\pi_j z_{nj}-\pi_k z_{nk}}{\sum_j\pi_j z_{nj}}\cdot(t_n-y_{nk})\phi_n^T+\gamma_{nk}y_{nk}(1-y_{nk})\cdot\phi_n\phi_n^T\right] \\
&= \sum_{n=1}^{N}\left[(y_{nk}-t_n)(t_n-y_{nk})\gamma_{nk}(1-\gamma_{nk})+\gamma_{nk}y_{nk}(1-y_{nk})\right]\phi_n\phi_n^T \\
&= \sum_{n=1}^{N}\gamma_{nk}\left[y_{nk}(1-y_{nk})-(1-\gamma_{nk})(t_n-y_{nk})^2\right]\phi_n\phi_n^T
\end{aligned}
$$

1.2 Problem 2

1.2.1 Prove that for each threshold s, there is some $m_0(s)\in\{0,1,\dots,m\}$ satisfying (1.16) and (1.17)

证明:

由于:

$$x^{(1)}>x^{(2)}>...>x^{(m)}$$

因此对于任意一个 s，存在一个 $m_0(s)\in\{0,1,\dots,m\}$，满足:

$$x^{(1)}>...>x^{(m_0(s))}\geq s\geq x^{(m_0(s)+1)}...>x^{(m)}$$

此时，对于 $\phi_{s,+}$，可得:

$$
\begin{aligned}
&\sum_{i=1}^{m}p_i\mathbb{I}\left\{y^{(i)}\neq\phi_{s,+}\left(x^{(i)}\right)\right\} \\
=&\sum_{i=1}^{m_s(0)}p_i\mathbb{I}\left\{y^{(i)}\neq 1\right\}+\sum_{i=m_s(0)+1}^{m}p_i\mathbb{I}\left\{y^{(i)}\neq -1\right\} \\
=&\sum_{i=1}^{m_s(0)}p_i\mathbb{I}\left\{y^{(i)}=-1\right\}+\sum_{i=m_s(0)+1}^{m}p_i\mathbb{I}\left\{y^{(i)}=1\right\} \\
=&\sum_{i=1}^{m_s(0)}p_i\frac{1-y^{(i)}}{2}+\sum_{i=m_s(0)+1}^{m}p_i\frac{1+y^{(i)}}{2} \\
=&\frac{1}{2}\sum_{i=1}^{m}p_i-\frac{1}{2}\left(\sum_{i=1}^{m_s(0)}y^{(i)}p_i-\sum_{i=m_s(0)+1}^{m}y^{(i)}p_i\right) \\
=&\frac{1}{2}-\frac{1}{2}\left(\sum_{i=1}^{m_s(0)}y^{(i)}p_i-\sum_{i=m_s(0)+1}^{m}y^{(i)}p_i\right)
\end{aligned}
$$

同理，由于 $\phi_{s,+}(x)=-\phi_{s,-}(x)$，易得:

$$\sum_{i=1}^{m}p_i\mathbb{I}\left\{y^{(i)}\neq\phi_{s,-}\left(x^{(i)}\right)\right\}=\frac{1}{2}-\frac{1}{2}\left(\sum_{i=m_s(0)+1}^{m}y^{(i)}p_i-\sum_{i=1}^{m_s(0)}y^{(i)}p_i\right)$$

对于 $m_0(s)=0$ 或 $m_0(s)=m$ 的特殊情况，对于空集求和的结果视为 0，则上式仍然满足，由此得证。

1.2.2 Define $f(m_0)$, prove that given $\gamma=\frac{1}{2m}$, $\max_{m_0}|f(m_0)|\geq 2\gamma$

证明:

由于:

$$f(m_0)=\sum_{i=1}^{m_0}y^{(i)}p_i-\sum_{i=m_0+1}^{m}y^{(i)}p_i$$

则对于 $0\leq m_0\leq m$:

$$\begin{aligned}|f(m_0)-f(m_0+1)|&=|\sum_{i=1}^{m_0}y^{(i)}p_i-\sum_{i=m_0+1}^{m}y^{(i)}p_i|\\&=|-2y^{(m_0+1)}p_{m_0+1}|\\&=2p_{m_0+1}\end{aligned}$$

假设，$\forall m_0\in\{0,1,...,m-1\}$，$p_{m_0+1}<\frac{1}{m}$，则:

$$\sum_{m_0=0}^{m-1}p_{m_0+1}=\sum_{i=1}^{m}p_i<m\cdot\frac{1}{m}=1$$

这与 $\sum_{i=1}^{m}p_i=1$ 矛盾，因此，$\exists m_0\in\{0,1,...,m-1\}$，$p_{m_0+1}\geq\frac{1}{m}$。

由此可得:

$$\max_{0\leq m_0\leq m-1}|f(m_0)-f(m_0+1)|\geq\frac{2}{m}$$

因此，可知:

$$2\cdot\max_{m_0}|f(m_0)|\geq\max_{0\leq m_0\leq m-1}|f(m_0)-f(m_0+1)|\geq\frac{2}{m}$$

即:

$$\max_{m_0}|f(m_0)|\geq 2\gamma=\frac{1}{m}$$

由此得证。

1.2.3 Give an upperbound on the number of thresholded decision stumps required to achieve zero error on a given training set.

给定训练集 $\{(x^{(i)},y^{(i)})\}_{i=1}^{m}$，不妨假设 $x^{(1)}\geq x^{(2)}\geq...\geq x^{(m)}$（可以通过重新排列使其满足）。由 (1.2.2) 可知，$\exists m_0\in\{0,1,...,m\}$，使得 $|f(m_0)|\geq\frac{1}{m}$，此时，选取 $x^{(m_0)}\leq s\leq x^{(m_0+1)}$，则由 (1.2.1) 可知:

$$\begin{aligned}\sum_{i=1}^{m}p_i\mathbb{I}\left\{y^{(i)}\neq\phi_{s,+}\left(x^{(i)}\right)\right\}&=\frac{1}{2}-\frac{1}{2}f(m_0)\\\sum_{i=1}^{m}p_i\mathbb{I}\left\{y^{(i)}\neq\phi_{s,-}\left(x^{(i)}\right)\right\}&=\frac{1}{2}+\frac{1}{2}f(m_0)\end{aligned}$$

由于 $f(m_0)\geq\frac{1}{m}$ 或 $f(m_0)\leq-\frac{1}{m}$，因此:

$$\begin{aligned}\sum_{i=1}^{m}p_i\mathbb{I}\left\{y^{(i)}\neq\phi_{s,+}\left(x^{(i)}\right)\right\}&=\frac{1}{2}-\frac{1}{2}f(m_0)\leq\frac{1}{2}-\frac{1}{2m}\\\sum_{i=1}^{m}p_i\mathbb{I}\left\{y^{(i)}\neq\phi_{s,-}\left(x^{(i)}\right)\right\}&=\frac{1}{2}+\frac{1}{2}f(m_0)\leq\frac{1}{2}-\frac{1}{2m}\end{aligned}$$

至少有一个成立，则我们可以构造出对应的弱分类器。此时分类器的 margin 为 $\gamma=\frac{1}{2m}$。

已知每一轮迭代，都可以构造一个分类器使得:

$$J_t \le \sqrt{1-4\gamma^2}J_{t-1}$$

由于 $J_0 = \frac{1}{2} - \gamma$，则要达到 zero error，需要使 $J_t < \frac{1}{m}$，即:

$$J_t \le \left(\frac{1}{2} - \frac{1}{2m}\right)\left(\sqrt{1-\frac{1}{m^2}}\right)^t \le \frac{1}{m}$$

只需要保证:

$$t \ge 2 \cdot \log_{\frac{m^2-1}{m^2}} \frac{2}{m-1} = t_{max}$$

因此，可以得到迭代轮次的上限为 t_{max}，即最多需要 $t_{max}+1$ 个弱分类器。

2 Deep Neural Networks

2.1 Problem 3

2.1.1 $\frac{\partial f_{\text{CE}}}{\partial b^{(2)}}$

首先计算:

$$\frac{\partial \hat{y}_{i,p}}{\partial z_{2,i,q}} = \begin{cases} \dfrac{e^{z_{2,i,p}}}{\sum e^z}(1 - \dfrac{e^{z_{2,i,p}}}{\sum e^z}) = \hat{y}_{i,p}(1-\hat{y}_{i,p}), & p = q \\ -\dfrac{e^{z_{2,i,p}}}{\sum e^z}\dfrac{e^{z_{2,i,q}}}{\sum e^z} = -\hat{y}_{i,p}\hat{y}_{i,q}, & p \ne q \end{cases}$$

由此可得:

$$\begin{aligned} \frac{\partial f_{\text{CE}}}{\partial z_{2,i,p}} &= -\frac{1}{n}\sum_{q=1}^{n_y} y_{i,q}\frac{1}{\hat{y}_{i,q}}\frac{\partial \hat{y}_{i,q}}{\partial z_{2,i,p}} \\ &= -\frac{1}{n}(\sum_{q\ne p} -y_{i,q}\hat{y}_{i,p} + y_{i,p}(1-\hat{y}_{i,p})) \\ &= \frac{1}{n}(\hat{y}_{i,p}\sum_q y_{i,q} - y_{i,p}) \end{aligned}$$

进一步地，可得:

$$\begin{aligned} \frac{\partial f_{\text{CE}}}{\partial b_p^{(2)}} &= \sum_{i=1}^{n}\sum_{q=1}^{n_y}\frac{\partial f_{\text{CE}}}{\partial z_{2,i,q}} \cdot \frac{\partial z_{2,i,q}}{\partial b_p^{(2)}} \\ &= \sum_{i=1}^{n}\frac{\partial f_{\text{CE}}}{\partial z_{2,i,p}} \cdot 1 \\ &= \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_{i,p}\sum_q y_{i,q} - y_{i,p}) \end{aligned}$$

整理可得:

$$\frac{\partial f_{\text{CE}}}{\partial \boldsymbol{b}^{(2)}} = \frac{1}{n}\sum_{i=1}^{n}((\boldsymbol{y}_i^T \cdot \mathbf{1})\hat{\boldsymbol{y}}_i - \boldsymbol{y}_i)$$

2.1.2 $\frac{\partial f_{\text{CE}}}{\partial W^{(2)}}$

首先计算:

$$\frac{\partial z_{2,i,k}}{\partial W_{p,q}^{(2)}} = \hat{h}_{1,i,q}$$

由此可得:

$$
\begin{aligned}
\frac{\partial f_{\mathrm{CE}}}{\partial W_{p,q}^{(2)}} &= \sum_{i=1}^{n}\sum_{k=1}^{n_y}\frac{\partial f_{\mathrm{CE}}}{\partial z_{2,i,k}}\cdot\frac{\partial z_{2,i,k}}{\partial W_{p,q}^{(2)}} \\
&= \sum_{i=1}^{n}\frac{1}{n}(\hat{y}_{i,p}\sum_{k}y_{i,k}-y_{i,p})\cdot\hat{h}_{1,i,q} \\
&= \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_{i,p}\sum_{k}y_{i,k}-y_{i,p})\hat{h}_{1,i,q}
\end{aligned}
$$

整理可得:

$$
\frac{\partial f_{\mathrm{CE}}}{\partial \boldsymbol{W}^{(2)}} = \frac{1}{n}\sum_{i=1}^{n}((\boldsymbol{y}_i^T\cdot\mathbf{1})\hat{\boldsymbol{y}}_i-\boldsymbol{y}_i)\cdot\hat{\boldsymbol{h}}_{1,i}^T
$$

2.1.3 $\frac{\partial f_{\mathrm{CE}}}{\partial \beta}$

首先计算:

$$
\begin{aligned}
\frac{\partial z_{2,i,p}}{\partial \beta} &= \sum_{k=1}^{n_y}\frac{\partial z_{2,i,p}}{\partial \hat{h}_{1,i,k}}\cdot\frac{\partial \hat{h}_{1,i,k}}{\partial \beta} \\
&= \sum_{k=1}^{n_1}W_{p,k}^{(2)}\cdot 1
\end{aligned}
$$

由此可得:

$$
\begin{aligned}
\frac{\partial f_{\mathrm{CE}}}{\partial \beta} &= \sum_{i=1}^{n}\sum_{p=1}^{n_y}\frac{\partial f_{\mathrm{CE}}}{\partial z_{2,i,p}}\cdot\frac{\partial z_{2,i,p}}{\partial \beta} \\
&= \sum_{i=1}^{n}\sum_{p=1}^{n_y}\frac{1}{n}(\hat{y}_{i,p}\sum_{q}y_{i,q}-y_{i,p})\cdot\sum_{k=1}^{n_1}W_{p,k}^{(2)} \\
&= \frac{1}{n}\sum_{i=1}^{n}((\boldsymbol{y}_i^T\cdot\mathbf{1})\hat{\boldsymbol{y}}_i-\boldsymbol{y}_i)^T\cdot(\boldsymbol{W}^{(2)}\cdot\mathbf{1}_{n_1\times 1})
\end{aligned}
$$

2.1.4 $\frac{\partial f_{\mathrm{CE}}}{\partial \gamma}$

首先计算:

$$
\begin{aligned}
\frac{\partial z_{2,i,p}}{\partial \gamma} &= \sum_{k=1}^{n_1}\frac{\partial z_{2,i,p}}{\partial \hat{h}_{1,i,k}}\cdot\frac{\partial \hat{h}_{1,i,k}}{\partial \gamma} \\
&= \sum_{k=1}^{n_1}W_{p,k}^{(2)}\frac{h_{1,i,k}-\mu_k}{\sqrt{\sigma_k^2+\epsilon}}
\end{aligned}
$$

由此可得:

$$
\begin{aligned}
\frac{\partial f_{\mathrm{CE}}}{\partial \gamma} &= \sum_{i=1}^{n}\sum_{p=1}^{n_y}\frac{\partial f_{\mathrm{CE}}}{\partial z_{2,i,p}}\cdot\frac{\partial z_{2,i,p}}{\partial \gamma} \\
&= \sum_{i=1}^{n}\sum_{p=1}^{n_y}\frac{1}{n}(\hat{y}_{i,p}\sum_{q}y_{i,q}-y_{i,p})\cdot\sum_{k=1}^{n_1}W_{p,k}^{(2)}\frac{h_{1,i,k}-\mu_k}{\sqrt{\sigma_k^2+\epsilon}} \\
&= \frac{1}{n}\sum_{i=1}^{n}((\boldsymbol{y}_i^T\cdot\mathbf{1})\hat{\boldsymbol{y}}_i-\boldsymbol{y}_i)^T\cdot(\boldsymbol{W}^{(2)}\cdot\left[\frac{h_{1,i,1}-\mu_1}{\sqrt{\sigma_1^2+\epsilon}},\dots,\frac{h_{1,i,n_1}-\mu_{n_1}}{\sqrt{\sigma_{n_1}^2+\epsilon}}\right]^T)
\end{aligned}
$$

令 $\hat{\boldsymbol{\sigma}} = \left(\frac{1}{\sqrt{\sigma_1^2+\epsilon}},\dots,\frac{1}{\sqrt{\sigma_{n_1}^2+\epsilon}}\right)^T$，则:

$$
\frac{\partial f_{\mathrm{CE}}}{\partial \gamma} = \frac{1}{n}\sum_{i=1}^{n}((\boldsymbol{y}_i^T\cdot\mathbf{1})\hat{\boldsymbol{y}}_i-\boldsymbol{y}_i)^T\cdot(\boldsymbol{W}^{(2)}\cdot((\boldsymbol{h}_{1,i}-\boldsymbol{\mu})\odot\hat{\boldsymbol{\sigma}}))
$$

其中，$\odot$ 表示向量对应元素相乘。

2.1.5 $\frac{\partial f_{\mathrm{CE}}}{\partial b^{(1)}}$

首先计算:

$$
\begin{aligned}
\frac{\partial f_{\mathrm{CE}}}{\partial h_{1,i,k}} &= \sum_{p=1}^{n_1}\left(\frac{\partial f_{\mathrm{CE}}}{\partial \hat{h}_{1,i,p}} \cdot \frac{\partial \hat{h}_{1,i,p}}{\partial h_{1,i,k}} + \frac{\partial f_{\mathrm{CE}}}{\partial \mu_p} \cdot \frac{\partial \mu_p}{\partial h_{1,i,k}} + \frac{\partial f_{\mathrm{CE}}}{\partial \sigma_p^2} \cdot \frac{\partial \sigma_p^2}{\partial h_{1,i,k}}\right) \\
&= \frac{\partial f_{\mathrm{CE}}}{\partial \hat{h}_{1,i,k}} \cdot \frac{\partial \hat{h}_{1,i,k}}{\partial h_{1,i,k}} + \frac{\partial f_{\mathrm{CE}}}{\partial \mu_k} \cdot \frac{\partial \mu_k}{\partial h_{1,i,k}} + \frac{\partial f_{\mathrm{CE}}}{\partial \sigma_k^2} \cdot \frac{\partial \sigma_k^2}{\partial h_{1,i,k}} \\
&= \frac{\partial f_{\mathrm{CE}}}{\partial \hat{h}_{1,i,k}} \cdot \frac{\partial \hat{h}_{1,i,k}}{\partial h_{1,i,k}} + \sum_{j=1}^{n} \frac{\partial f_{\mathrm{CE}}}{\partial \hat{h}_{1,j,k}} \cdot \left(\frac{\partial \hat{h}_{1,j,k}}{\partial \mu_k} + \frac{\partial \hat{h}_{1,j,k}}{\partial \sigma_k} \cdot \frac{\partial \sigma_k}{\partial \mu_k}\right) \cdot \frac{\partial \mu_k}{\partial h_{1,i,k}} + \frac{\partial f_{\mathrm{CE}}}{\partial \sigma_k^2} \cdot \frac{\partial \sigma_k^2}{\partial h_{1,i,k}} \\
&= \frac{\partial f_{\mathrm{CE}}}{\partial \hat{h}_{1,i,k}} \cdot \frac{\gamma}{\sqrt{\sigma_k^2+\epsilon}} + \sum_{j=1}^{n} \frac{\partial f_{\mathrm{CE}}}{\partial \hat{h}_{1,j,k}} \cdot \left(\frac{-\gamma}{\sqrt{\sigma_k^2+\epsilon}} + \frac{-\gamma(h_{1,j,k}-\mu_k)}{2(\sigma_k^2+\epsilon)\sqrt{\sigma_k^2+\epsilon}} \cdot \frac{-2}{n}\sum_{t=1}^{n}(h_{1,t,k}-\mu_k)\right) \cdot \frac{1}{n} \\
&\quad + \left(\sum_{j=1}^{n} \frac{\partial f_{\mathrm{CE}}}{\partial \hat{h}_{1,j,k}} \cdot \frac{-\gamma(h_{1,j,k}-\mu_k)}{2(\sigma_k^2+\epsilon)\sqrt{\sigma_k^2+\epsilon}}\right) \cdot \frac{2(h_{1,i,k}-\mu_k)}{n} \\
&= \frac{\gamma}{n\sqrt{\sigma_k^2+\epsilon}} \cdot \left[n\frac{\partial f_{\mathrm{CE}}}{\partial \hat{h}_{1,i,k}} - \sum_{j=1}^{n} \frac{\partial f_{\mathrm{CE}}}{\partial \hat{h}_{1,j,k}} - \frac{h_{1,i,k}-\mu_k}{\sqrt{\sigma_k^2+\epsilon}} \sum_{j=1}^{n} \frac{\partial f_{\mathrm{CE}}}{\partial \hat{h}_{1,j,k}} \cdot \frac{h_{1,j,k}-\mu_k}{\sqrt{\sigma_k^2+\epsilon}}\right]
\end{aligned}
$$

令 $\tilde{h}_{1,i,k} = \frac{h_{1,i,k}-\mu_k}{\sqrt{\sigma_k^2+\epsilon}}$，则可化简为:

$$
\begin{aligned}
\frac{\partial f_{\mathrm{CE}}}{\partial h_{1,i,k}} &= \frac{\gamma}{n\sqrt{\sigma_k^2+\epsilon}} \cdot \left[n\frac{\partial f_{\mathrm{CE}}}{\partial \hat{h}_{1,i,k}} - \sum_{j=1}^{n} \frac{\partial f_{\mathrm{CE}}}{\partial \hat{h}_{1,j,k}} - \tilde{h}_{1,i,k} \sum_{j=1}^{n} \frac{\partial f_{\mathrm{CE}}}{\partial \hat{h}_{1,j,k}} \cdot \tilde{h}_{1,j,k}\right] \\
&= \frac{\gamma}{n\sqrt{\sigma_k^2+\epsilon}} \cdot \left[n\frac{\partial f_{\mathrm{CE}}}{\partial \hat{h}_{1,i,k}} - \sum_{j=1}^{n} \frac{\partial f_{\mathrm{CE}}}{\partial \hat{h}_{1,j,k}}(1+\tilde{h}_{1,i,k}\tilde{h}_{1,j,k})\right]
\end{aligned}
$$

由于:

$$
\begin{aligned}
\frac{\partial h_{1,i,p}}{\partial b_q^{(1)}} &= \sum_{k=1}^{n_1} \frac{\partial h_{1,i,p}}{\partial z_{1,i,k}} \cdot \frac{\partial z_{1,i,k}}{\partial b_q^{(1)}} \\
&= \begin{cases} \dfrac{\partial h_{1,i,p}}{\partial z_{1,i,p}} \cdot 1, & p=q \\ 0, & p \neq q \end{cases} \\
&= \begin{cases} 1, & p=q,\ z_{1,i,p}>0 \\ 0, & \text{other} \end{cases}
\end{aligned}
$$

由此可得:

$$
\begin{aligned}
\frac{\partial f_{\mathrm{CE}}}{\partial b_k^{(1)}} &= \sum_{i=1}^{n} \frac{\partial f_{\mathrm{CE}}}{\partial h_{1,i,k}} \cdot \mathbb{I}\{z_{1,i,k}>0\} \\
&= \sum_{i=1}^{n} \frac{\gamma}{n\sqrt{\sigma_k^2+\epsilon}} \cdot \left[n\frac{\partial f_{\mathrm{CE}}}{\partial \hat{h}_{1,i,k}} - \sum_{j=1}^{n} \frac{\partial f_{\mathrm{CE}}}{\partial \hat{h}_{1,j,k}}(1+\tilde{h}_{1,i,k}\tilde{h}_{1,j,k})\right] \cdot \mathbb{I}\{z_{1,i,k}>0\} \\
&= \frac{\gamma}{n\sqrt{\sigma_k^2+\epsilon}} \sum_{i=1}^{n} \left[\sum_{p=1}^{n_1}(\hat{y}_{i,p}\sum_{q} y_{i,q} - y_{i,p}) \cdot W_{p,k}^{(2)} - \sum_{j=1}^{n}\left(\frac{1}{n}\sum_{p=1}^{n_1}(\hat{y}_{j,p}\sum_{q} y_{j,q} - y_{j,p}) \cdot W_{p,k}^{(2)}\right) \cdot (1+\tilde{h}_{1,i,k}\tilde{h}_{1,j,k})\right] \cdot \mathbb{I}\{z_{1,i,k}>0\}
\end{aligned}
$$

2.1.6 $\frac{\partial f_{\mathrm{CE}}}{\partial W^{(1)}}$

首先计算:

$$\begin{aligned}\frac{\partial h_{1,i,k}}{\partial W^{(1)}_{p,q}} &= \sum_{t=1}^{n_1} \frac{\partial h_{1,i,k}}{\partial z_{1,i,t}} \cdot \frac{\partial z_{1,i,t}}{\partial W^{(1)}_{p,q}} \\ &= \begin{cases} x_{i,q}, \quad k=p, \quad z_{1,i,k}>0 \\ 0, \quad \text{other} \end{cases}\end{aligned}$$

由此可得:

$$\begin{aligned}\frac{\partial f_{\text{CE}}}{\partial W^{(1)}_{p,q}} &= \sum_{i=1}^{n}\sum_{k=1}^{n_1} \frac{\partial f_{\text{CE}}}{\partial h_{1,i,k}} \cdot \frac{\partial h_{1,i,k}}{\partial W^{(1)}_{p,q}} \\ &= \sum_{i=1}^{n} \frac{\partial f_{\text{CE}}}{\partial h_{1,i,p}} \cdot x_{i,q} \cdot \mathbb{I}\{z_{1,i,p}>0\} \\ &= \frac{\gamma}{n\sqrt{\sigma_k^2+\epsilon}} \sum_{i=1}^{n} \left[\sum_{t=1}^{n_1} (\hat{y}_{i,t} \sum_r y_{i,r} - y_{i,t}) \cdot W^{(2)}_{t,p} - \sum_{j=1}^{n} \left(\frac{1}{n} \sum_{t=1}^{n_1} (\hat{y}_{j,t} \sum_r y_{j,r} - y_{j,t}) \cdot W^{(2)}_{t,p} \right) \cdot (1+\tilde{h}_{1,i,p}\tilde{h}_{1,j,p}) \right] \\ &\quad \cdot x_{i,q} \cdot \mathbb{I}\{z_{1,i,p}>0\}\end{aligned}$$

3 Clustering

3.1 Problem 4

模型的隐变量为每个文档对应的 topic，即 $c_d \in \{1,2,...,K\}$，需要求解参数 $\theta = \{\pi, \mu\}$，首先随机初始化得到初始值 $\theta^{(0)} = \{\pi^{(0)}, \mu^{(0)}\}$。以下进行 EM 算法推导。

3.1.1 E 步

给定上一轮迭代得到的参数 $\theta^{(l-1)} = \{\pi^{(l-1)}, \mu^{(l-1)}\}$，由于各个文档是独立分布的，因此隐变量的条件概率为:

$$\begin{aligned}P(C|T,\theta^{(l-1)}) &= \prod_{d=1}^{D} P(c_d=k|d,\theta^{(l-1)}) \\ &= \prod_{d=1}^{D} \frac{P(c_d=k)\cdot P(d|c_d=k)}{P(d)} \\ &= \prod_{d=1}^{D} \frac{\pi_k^{(l-1)} \cdot \frac{n_d!}{\prod_w T_{dw}!} \prod_w {\mu_{wk}^{(l-1)}}^{T_{dw}}}{\frac{n_d!}{\prod_w T_{dw}!} \sum_{k=1}^{K} \pi_k^{(l-1)} \prod_w {\mu_{wk}^{(l-1)}}^{T_{dw}}} \\ &= \prod_{d=1}^{D} \frac{\pi_k^{(l-1)} \cdot \prod_w {\mu_{wk}^{(l-1)}}^{T_{dw}}}{\sum_{k=1}^{K} \pi_k^{(l-1)} \prod_w {\mu_{wk}^{(l-1)}}^{T_{dw}}}\end{aligned}$$

令责任 $r_{dk}^{(l)} = \dfrac{\pi_k^{(l-1)} \cdot \prod\limits_w {\mu_{wk}^{(l-1)}}^{T_{dw}}}{\sum\limits_{k=1}^{K} \pi_k^{(l-1)} \prod\limits_w {\mu_{wk}^{(l-1)}}^{T_{dw}}}$，则由此可以求得其似然的条件概率期望:

$$\begin{aligned}
Q(\theta, \theta^{(l-1)}) &= \sum_C P(C|T, \theta^{(l-1)}) \log P(C, T|\theta) \\
&= \sum_{k=1}^{K} \sum_{d=1}^{D} r_{dk}^{(l)} \log P(c_d = k) P(d|c_d = k, \theta^{(l-1)}) \\
&= \sum_{k=1}^{K} \sum_{d=1}^{D} r_{dk}^{(l)} \left(\log \pi_k + \log \frac{n_d!}{\prod_w T_{dw}!} \prod_w {\mu_{wk}}^{T_{dw}} \right)
\end{aligned}$$

3.1.2 M 步

最优化函数 $Q(\theta, \theta^{(l-1)})$ 去更新参数 θ。

由于 $\mu_k = (\mu_{1k}, \mu_{2k}, ...\mu_{Wk})$ 存在限制条件: $\sum_{w=1}^{W} \mu_{wk} = 1, \quad k \in \{1, 2, ..., K\}$，因此使用拉特朗日乘子法，对给定的一个 k，构造函数:

$$F(\mu_k, \lambda) = Q(\theta, \theta^{(l-1)}) + \lambda \left(\sum_{w=1}^{W} \mu_{wk} - 1 \right)$$

对 $\{\mu_{wk}\}_{w=1}^{W}$ 和 λ 分别求偏导，并令偏导为 0，可得:

$$\begin{cases}
\dfrac{\partial F}{\partial \mu_{wk}} = \sum\limits_{d=1}^{D} r_{dk}^{(l)} \dfrac{T_{dw}}{\mu_{wk}} + \lambda = 0, \quad w = 1, 2, ..., W \\
\dfrac{\partial F}{\partial \lambda} = \sum\limits_{w=1}^{W} \mu_{wk} - 1 = 0
\end{cases}$$

求解可得:

$$\lambda = -\sum_{d=1}^{D} \sum_{w=1}^{W} r_{dk}^{(l)} \cdot T_{dw} = -\sum_{d=1}^{D} r_{dk}^{(l)} \cdot n_d$$

由此可得 μ_{wk} 的更新为:

$$\mu_{wk}^{(l)} = \frac{\sum_{d=1}^{D} r_{dk}^{(l)} \cdot T_{dw}}{\sum_{d=1}^{D} r_{dk}^{(l)} \cdot n_d}, \quad k = 1, 2, ..., K, \ w = 1, 2, ..., W$$

同理，由限制条件 $\sum_{k=1}^{K} \pi_k = 1$，构造函数:

$$F(\pi_k, \lambda) = Q(\theta, \theta^{(l-1)}) + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right)$$

对 $\{\pi_k\}_{k=1}^{K}$ 和 λ 分别求偏导，并令偏导为 0，可得:

$$\begin{cases}
\dfrac{\partial F}{\partial \pi_k} = \sum\limits_{d=1}^{D} \dfrac{r_{dk}^{(l)}}{\pi_k} + \lambda = 0, \quad k = 1, 2, ..., K \\
\dfrac{\partial F}{\partial \lambda} = \sum\limits_{k=1}^{K} \pi_k - 1 = 0
\end{cases}$$

求解可得:

$$\lambda = -\sum_{d=1}^{D} \sum_{k=1}^{K} r_{dk}^{(l)}$$

由此可得 π_k 的更新为:

$$\pi_k^{(l)} = \frac{\sum_{d=1}^{D} r_{dk}^{(l)}}{\sum_{d=1}^{D}\sum_{k=1}^{K} r_{dk}^{(l)}}, \quad k = 1, 2, ..., K$$

EM 算法不断重复以上两步，直到算法收敛或者达到某个最大迭代步数 L。

3.2 Problem 5

按照上述推导实现 EM 算法，并且设置收敛条件为:

$$||\theta^{(l)} - \theta^{(l-1)}||^2 \leq 1 \times 10^{-6}$$

或者达到最大迭代次数 $L = 10$。

设置参数 K 分别为 $10, 20, 30, 50$，得到每一个 topic 的分布 μ_k，根据 μ_k 选择概率最大的 Top p 个词 (对于 K=10，20 给出 Top 10，对于 K=30 给出 Top 5，对于 K=50 给出 Top 3)，结果如下:

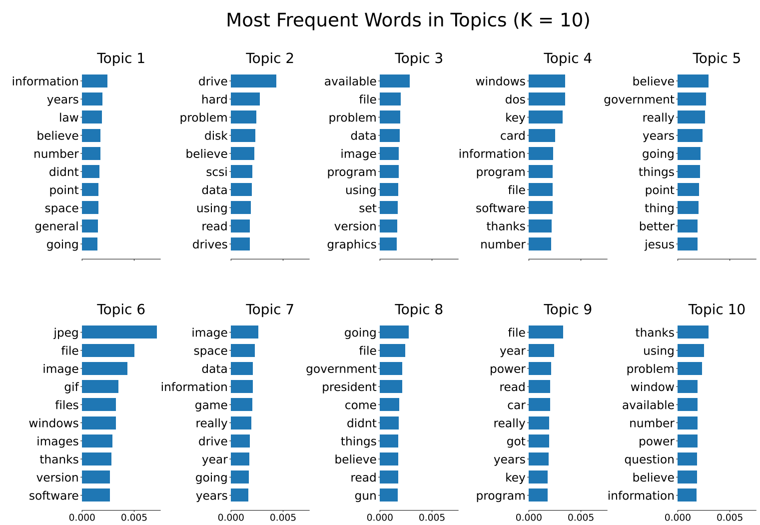

图 1: K=10 情况下高频词结果

3.2.1

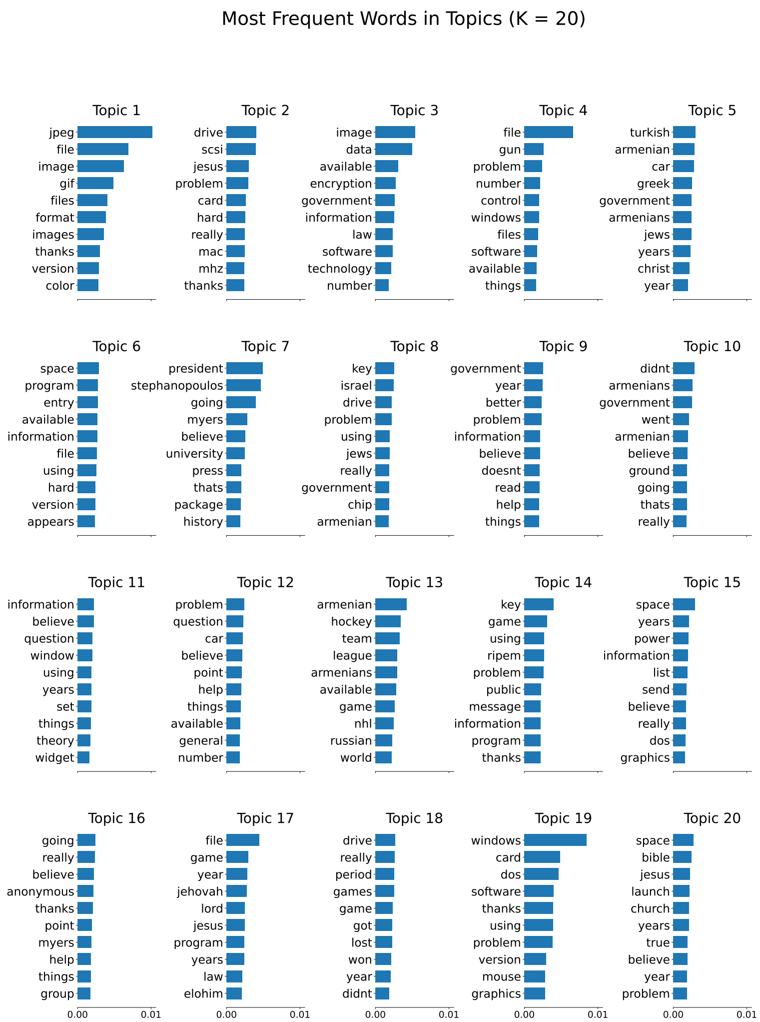

图 2: K=20 情况下高频词结果

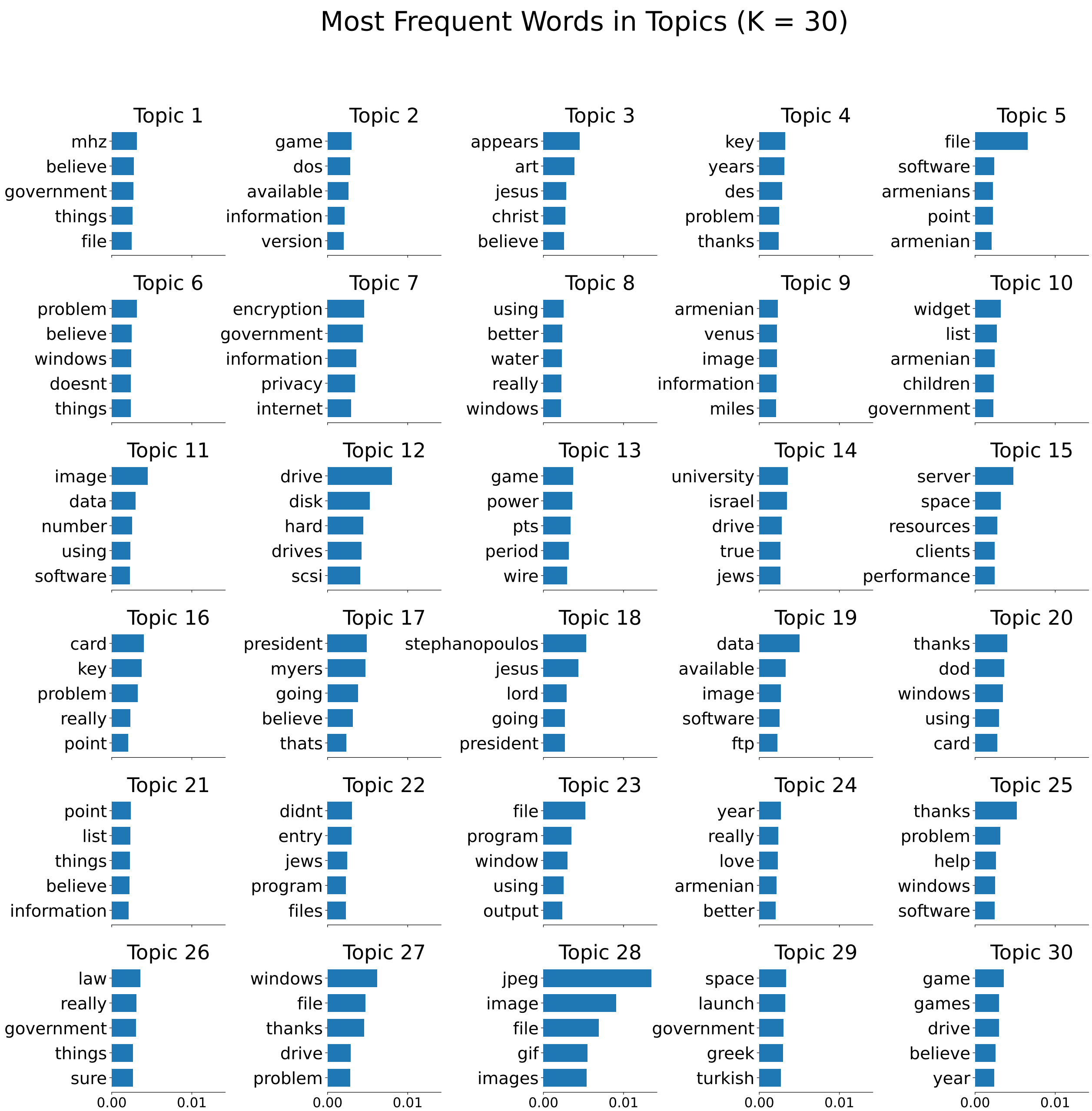

图 3: K=30 情况下高频词结果

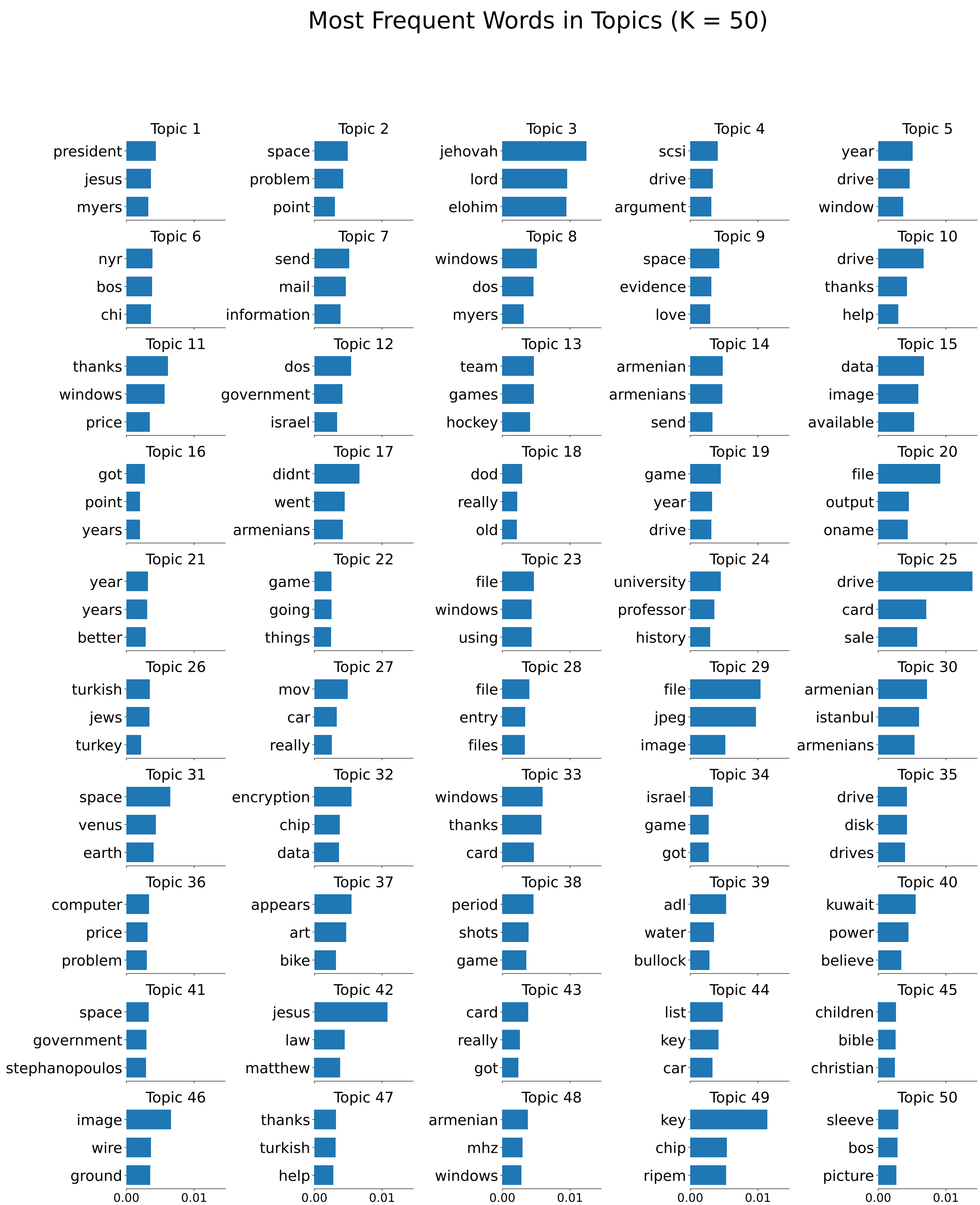

图 4: K=50 情况下高频词结果