Submodularity in Data Subset Selection and Active Learning

ICML 2015

Introducton

Select a subset of big data to train a classifier while incurring minimal performance loss.
We show the connection of submodularity to the data likelihood functions for Naive Bayes (NB) and Nearest Neighbor (NN) classifiers, and formulate the data subset selection problems for these classifiers as constrained submodular maximization.

Naive Bayes Classifier

$$p(C_k|\boldsymbol{x}) = \frac{p(C_k)p(\boldsymbol{x}|C_k)}{p(\boldsymbol{x})}$$

$$\bar{y} = \underset{k\in\{1,\ldots,K\}}{argmax}\, p(C_k)\prod_{i=1}^{d} p(x_i|C_k)$$

Naive Bayes Classifier

$$V = \{(x^i, y^i)\}_{i=1}^m \; S \subseteq V$$

$$\theta_{x_j|y} = p(x_j|y) \text{ and } \theta_y = p(y)$$

$$\theta_{x_j|y}(S) = \frac{m_{x_j,y}(S)}{m_y(S)}, \theta_y(S) = \frac{m_y(S)}{|S|}$$

$$m_{x_j,y}(S) = \sum_{i \in S} 1\{x_j^i = x_j \wedge y^i = y\}$$

$$m_y(S) = \sum_{i \in S} 1\{y^i = y\}$$

Naive Bayes Classifier

data log-likelihood set function:

$$\ell^{\mathrm{NB}}(S) = \sum_{i \in V} \log p(x^i, y^i; \theta(S))$$

Naive Bayes Classifier

$$\ell^{\mathrm{NB}}(S) = \underbrace{\sum_{j=1}^{d} \sum_{x_j \in \mathcal{X}} \sum_{y \in \mathcal{Y}} m_{x_j,y}(V) \log(m_{x_j,y}(S))}_{\text{term 1: } f_{\mathrm{NB}}(S)}$$

$$- \underbrace{(d-1) \sum_{y \in \mathcal{Y}} m_y(V) \log(m_y(S))}_{\text{term 2}} - \underbrace{|V| \log |S|}_{\text{term 3}}$$

$|S \cap V^y| = k \frac{|V^y|}{|V|}$

$|S| = k$

Naive Bayes Classifier

$$\max_{S \in \mathcal{B}(\mathcal{M})} f_{\mathrm{NB}}(S)$$

$$\mathcal{B}(\mathcal{M}) = \{S \subseteq V : |S \cap V^y| = k\frac{|V^y|}{|V|}, \forall y \in \mathcal{Y}\}$$

Naive Bayes Classifier

Theorem 1. *Let* $D_{KL}(p(x,y;\theta(V))||p(x,y;\theta(S))) \triangleq \sum_{x\in\mathcal{X}^d}\sum_{y\in\mathcal{Y}} p(x,y;\theta(V))\log\frac{p(x,y;\theta(V))}{p(x,y;\theta(S))}$ *be the KL-divergence between* $p(x,y;\theta(V))$ *and* $p(x,y;\theta(S))$, *where* $p(x,y;\theta(S))$ *is the maximum likelihood estimate of the joint distribution given a data set* S. *Under the Naïve Bayes assumption, Problem 1 is equivalent to*

$$\min_{|S|=k} D_{KL}(p(x,y;\theta(V))||p(x,y;\theta(S))).$$

Naive Bayes Classifier

$$\theta^{\alpha}_{x_j|y}(S) = \frac{m_{x_j,y}(S)+\alpha}{m_y(S)+\alpha|\mathcal{X}|} \qquad \theta^{\alpha}_{y}(S) = \frac{m_y(S)+\alpha}{|S|+\alpha|\mathcal{Y}|}$$

$$V' = V \cup \{v'\}$$

$$m_{x_j,y}(v') = \alpha, \forall x_j \in \mathcal{X}, y \in \mathcal{Y}, j = 1, \ldots, d$$

$$f'_{\mathrm{NB}(\alpha)}(S) = \sum_{j=1}^{d} \sum_{x_j \in \mathcal{X}} \sum_{y \in \mathcal{Y}} m_{x_j,y}(V) \log(m_{x_j,y}(S))$$

$$f_{\mathrm{NB}(\alpha)}(S) = f'_{\mathrm{NB}(\alpha)}(S \cup \{v'\}) - f'_{\mathrm{NB}(\alpha)}(\{v'\})$$

Nearest Neighbor Classification

$$\ell^{\mathrm{NN}}(S) = \sum_{i \in V} (\log p(x^i | y^i; \theta(S)) + \log p(y^i; \theta(S)))$$

$$\begin{aligned} p(x^i | y^i; \theta(S)) &= ce^{-\|x^i - x^j\|_2^2} = ce^{w(i,j)-d} \\ &= c' \exp(\max_{s \in S \cap V^{y^i}} w(i,s)) \end{aligned}$$

$$\log p(x^i | y^i; \theta(S)) = \log c' + \max_{s \in S \cap V^{y^i}} w(i,s)$$

Nearest Neighbor Classification

$$\ell^{\mathrm{NN}}(S) = \underbrace{\sum_{y \in \mathcal{Y}} \sum_{i \in V^y} \max_{s \in S \cap V^y} w(i,s)}_{\text{term 1: } f_{\mathrm{NN}}} + \underbrace{\sum_{y \in \mathcal{Y}} m_y(V) \log m_y(S)}_{\text{term 2}}$$

$$- \underbrace{|V| \log |S|}_{\text{term 3}} + \underbrace{C}_{\text{constant}} .$$

Active learning

Algorithm 1 Filtered Active Submodular Selection

Input: $\mathcal{U}$, T, B, $\{\beta_t\}_{t=1}^{T}$, Starting set of labels $\mathcal{L}$

for $t = 1, \cdots, T$ **do**

  Train the classifier using the labeled set $\mathcal{L}$, and derive the uncertainty scores δ^t;

  $\mathcal{U}^t \in \text{argmax}_{U \subseteq \mathcal{U} \setminus \mathcal{L}; |U| = \beta_t} \sum_{u \in U} \delta_u^t$;

  Obtain the most probable labels as the hypothesized labels $\{\hat{y}_u\}_{u \in \mathcal{U}^t}$.

  Instantiate $\hat{f}_t : 2^{\mathcal{U}^t} \to \mathbb{R}_+$ on $\{\hat{y}_u\}_{u \in \mathcal{U}^t}$ and $\mathcal{U}^t$;

  Find $L^t \in \text{argmax}_{|S| = B; S \subseteq \mathcal{U}^t} \hat{f}_t(S)$.

  $\mathcal{L} = \mathcal{L} \cup L^t$.

end for

Experiment

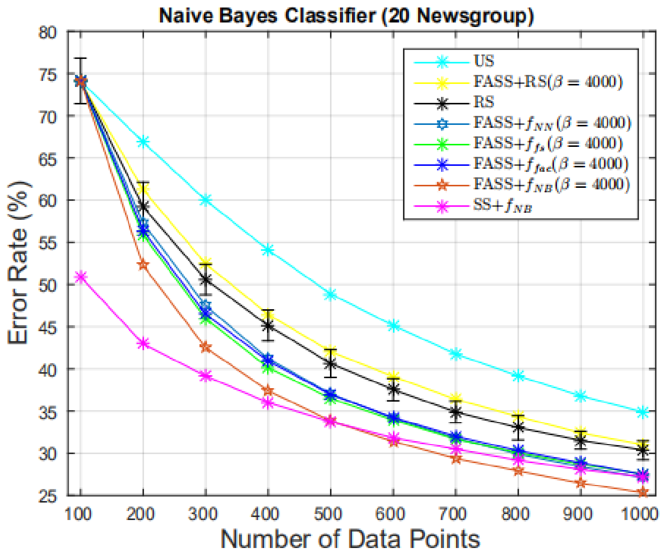

Figure 2.

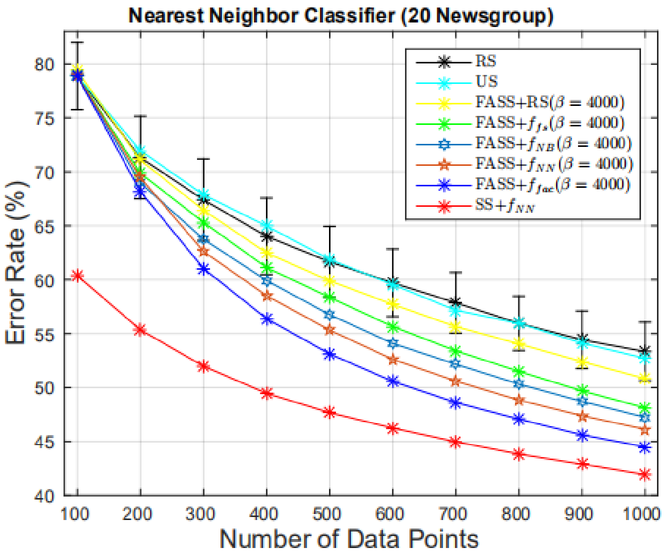

Figure 3.

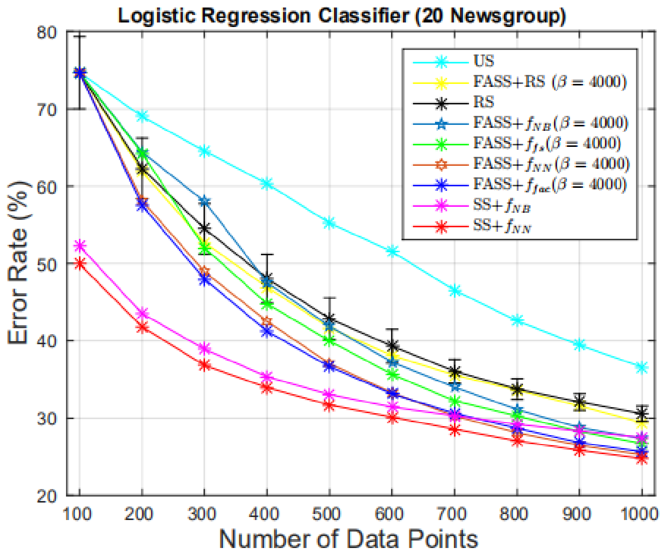

Figure 4.

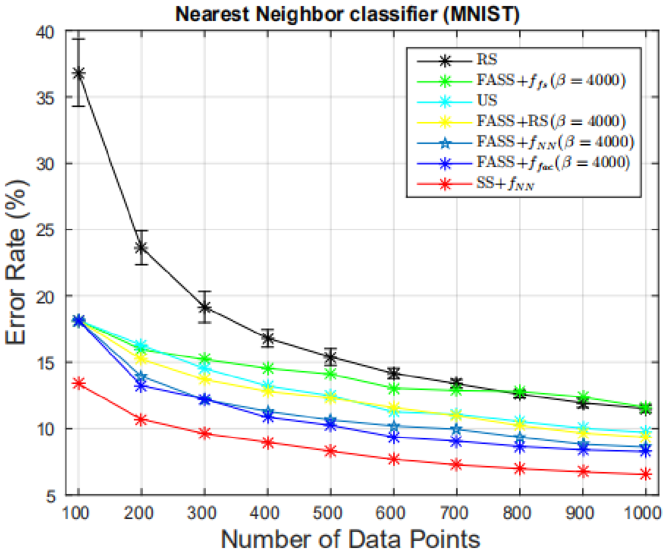

Figure 5.

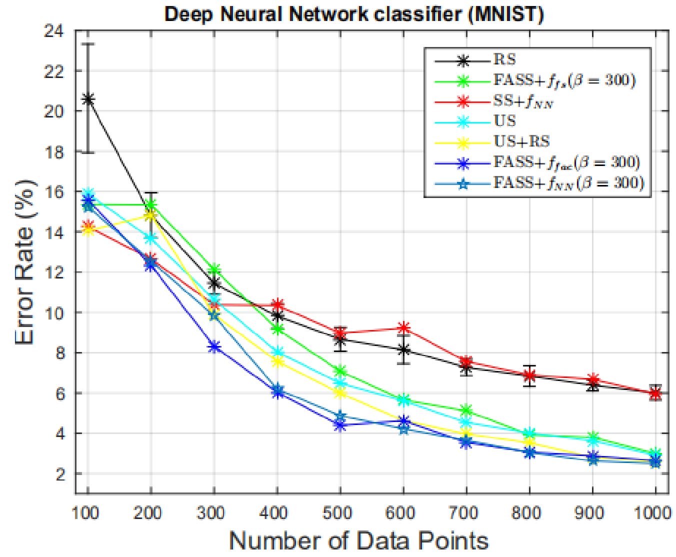

Figure 6.

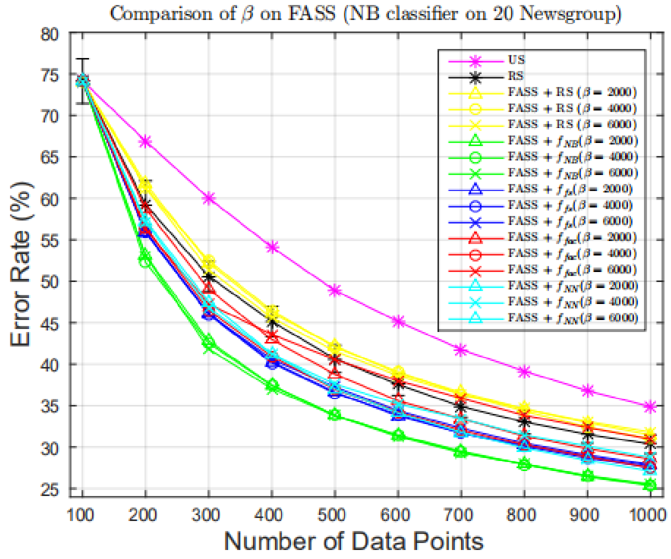

Figure 7.