Feature Selection using Stochastic Gates

Hengjun Jiang

Feature Selection

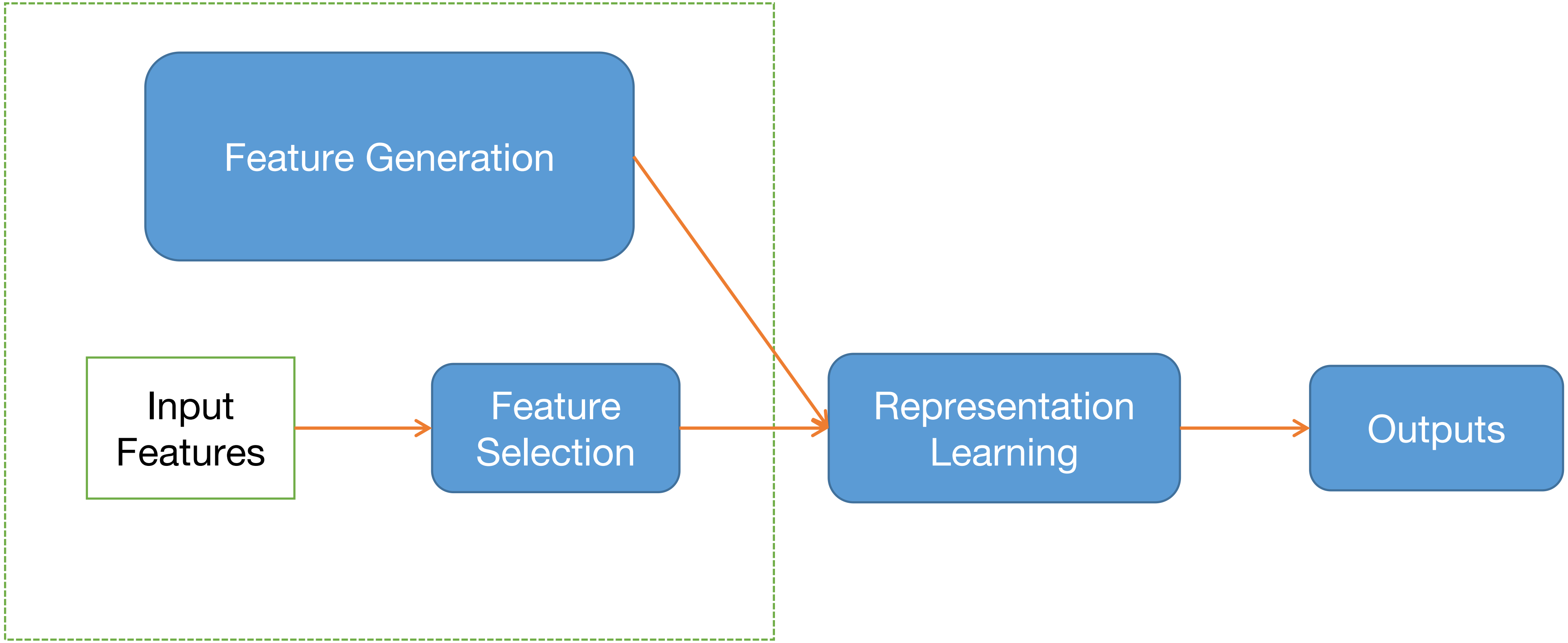
Feature Selection Definition:

Select a subset of meaningful(relevant) features from original features.

Why Feature Selection

- model interpretation
- memory cost
- convergence rate

Difference with Representation Learning



Feature Selection Methods

- Filter Methods
  - remove irrelevant features based on some pre-defined statistical measure scores(Mutual Information, Dependence Estimation, etc)
- Wrapper Methods
  - learn a subset of relevant features according to the output of a task-oriented model(Random Forests, Xgboost, etc)
- Embedded Methods
  - learn a task-oriented model while simultaneously selecting a subset of relevant features(LASSO(linear models), Neural Networks with L-p norm(nonlinear LASSO), etc)

Problem Formulation

Given n i.i.d. samples {(x_i, y_i), i=1, 2, ..., n}($\boldsymbol{x}_i \in R^d$, with features S of size d) generated from an unknown joint distribution $P_{X,Y}$, to select a subset of features $\mathcal{T}$ of size m(m<=d)

- dependence perspective(filter method):
  - select a subset of features $\mathcal{T}$ of size m and the remaining features $S \setminus \tau$ conditionally independent of Y given $\mathcal{T}$

- prediction perspective(wrapper method):
  - select a subset of features $\mathcal{T}$ and X_τ can well predict Y within a specific learning problem

$$\min_{\tau:|\tau|\leq m} \varepsilon_F(X_\tau) = \min_{\tau:|\tau|\leq m} \inf_{f\in F_m} E_{X,Y} L(Y, f(X_\tau))$$

Embeded Methods

LASSO(for linear models):

$$\min_{\boldsymbol{\theta}} \frac{1}{N} \sum_{n=1}^{N} L(\boldsymbol{\theta}^T \boldsymbol{x}_n, y_n) \;\; s.t. \parallel \boldsymbol{\theta} \parallel_1 \leq m.$$

Limitation: original LASSO optimization problem just shrink the coefficients $\boldsymbol{\theta}$ to approximately select the relevant features.

Embeded Methods

Given n i.i.d. samples {(x_i, y_i), i=1, 2, ..., N}($\boldsymbol{x}_i \in R^D,$ with features size D) generated from an unknown joint distribution $P_{X,Y}$, to select a subset of features X_S with feature indices $S \subset \{1, ..., D\}$, and simultaneously learn a model to predict Y based on the selected features X_S

$l_0\ norm$ LASSO(for nonlinear models):

$$\min_{\boldsymbol{\theta},\boldsymbol{s}} \frac{1}{N}\sum_{n=1}^{N} L(f_{\boldsymbol{\theta}}(\boldsymbol{x}_n \odot \boldsymbol{s}), y_n)\ \ s.t.\ \| \ \boldsymbol{s}\ \|_0 \leq m.$$

$$where\ \boldsymbol{s} = \{0,1\}^D$$

Probabilistic Perspective

let $\widetilde{S} \in R^D$ be a random vector sampled from a Multivariate Bernoulli distribution with i.i.d. entries, and $\pi_d = p(\widetilde{S}_d = 1),\ \forall d \in \{1, ..., D\}$

ERR(Empirical Regularized Risk):

$$\min_{\boldsymbol{\theta},\boldsymbol{\pi}} \widehat{R}(\boldsymbol{\theta}, \boldsymbol{\pi}) = \min_{\boldsymbol{\theta},\boldsymbol{\pi}} \widehat{E_{X,Y}} E_{\widetilde{S}}[L(f_{\boldsymbol{\theta}}(\boldsymbol{X} \odot \widetilde{S}), Y) + \lambda \parallel \widetilde{S} \parallel_0]$$

$$where\ \widetilde{S} = \{0, 1\}^D$$

Probabilistic Perspective

ERR(Empirical Regularized Risk):

$$\min_{\boldsymbol{\theta},\boldsymbol{\pi}} \widehat{R}(\boldsymbol{\theta},\boldsymbol{\pi}) = \min_{\boldsymbol{\theta},\boldsymbol{\pi}} \widehat{E_{X,Y}} E_{\widetilde{S}}[L(f_{\boldsymbol{\theta}}(\boldsymbol{X} \odot \widetilde{S}), Y) + \lambda \parallel \widetilde{S} \parallel_0]$$

$$where\ \widetilde{S} = \{0,1\}^D$$

The first term's gradients with respect to $\pi_d,\ \forall d \in \{1, ..., D\}$ suffer from high variance if be estimated with REINFORCE(Monte Carlo).

The second term's gradients with respect to π_d is differentiable, cause:

$$E_{\widetilde{S}}[\lambda \parallel \widetilde{S} \parallel_0] = \lambda \sum_{d=1}^{D} \pi_d$$

Reparametrization Trick

Objective: $\min_{\boldsymbol{\theta},\boldsymbol{\beta}} E_{x\sim G_{\boldsymbol{\beta}}}[L(f_{\boldsymbol{\theta}}(x), y)],\ where\ x\ is\ a\ discrete\ random\ variable(vector)$

Q:

How to efficiently back propagate the gradients $\bigtriangledown E_{x\sim G_{\boldsymbol{\beta}}}[L(f_{\boldsymbol{\theta}}(x), y)]$ to $G_{\boldsymbol{\beta}}$?

A:

- REINFORCE-Monte Carlo(simple but suffers from high variance)
- Reparametrization Trick(simple and have low variance estimation)
  - p.s. reparametrization trick also works when x is a continuous random variable(vector), e.g. VAEs.

Reparametrization Trick

How reparametrization trick works?

Reparametrization Trick

Consider a random variable(vector) : $\boldsymbol{x} \sim P_{\boldsymbol{X}},\ \boldsymbol{x} \in R^d$

$$with\ mean\ \boldsymbol{\mu}\ and\ a\ large\ diagonal\ entries\ of\ covariance\ matrix\ \boldsymbol{\sigma}^2$$

How to calculate the expectation of L(X): $E[L(X)] = \sum P(\boldsymbol{x}) * L(\boldsymbol{x})$ where: $L(\boldsymbol{x}) : R^d \rightarrow R$

(Only)Monte Carlo:

$$E[L(X)] = \sum P(\boldsymbol{x}) * L(\boldsymbol{x}) \approx \frac{1}{K} \sum_{k=1}^{K} L(\boldsymbol{x}_k),\ where\ \boldsymbol{x}_k \sim P_{\boldsymbol{X}}$$

Chebyshev's Inequality:

$$P(|\bar{X} - \mu| > \epsilon) \leq \frac{D(\bar{X})}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$$

We can conclude:

when sample size n is small and variance σ^2 is large, it's a high variance estimation!

Reparametrization Trick

Consider a random variable(vector) : $\boldsymbol{x} \sim P_{\mathbf{X}}, \ \boldsymbol{x} \in R^d$

$$with\ mean\ \boldsymbol{\mu}\ and\ a\ large\ diagonal\ entries\ of\ covariance\ matrix\ \boldsymbol{\sigma}^2$$

How to calculate the expectation of L(X): $E[L(X)] = \sum P(\boldsymbol{x}) * L(\boldsymbol{x})$ where: $L(\boldsymbol{x}) : R^d \to R$

Reparametrization Trick:

$$let\ \boldsymbol{z} = \boldsymbol{\mu} + \boldsymbol{\sigma}\boldsymbol{\epsilon},\ where\ \boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{\sigma}_\epsilon^2)\ and\ \sigma_\epsilon^2 < \sigma^2$$

$$E[L(X)] \approx E_Z[L(Z)] = \sum P(\boldsymbol{z}) * L(\boldsymbol{z}) = \sum P(\boldsymbol{\epsilon}) * L(\boldsymbol{z}) \approx \frac{1}{K} \sum_{k=1}^{K} L(\boldsymbol{\mu} + \boldsymbol{\sigma}\boldsymbol{\epsilon}_k), where \boldsymbol{\epsilon}_k \sim N(\mathbf{0}, \boldsymbol{\sigma}_\epsilon^2)$$

Chebyshev's Inequality:

$$P(|\bar{X} - \mu| > \epsilon) \leq \frac{D(\bar{X})}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$$

Cause $\sigma_\epsilon^2 < \sigma^2$, so the estimation has lower variance, and also, it's fully differentiable.

Reparametrization Trick

$$let\ z_d = g(\mu_d + \epsilon_d) = max(0,\ min(1,\ \sigma_d \epsilon_d + \mu_d + 0.5))$$

$$where\ \epsilon_d\ sampled\ from\ a\ Gaussian\ N(0,\ \sigma_d^2)$$

$$\min_{\boldsymbol{\theta},\boldsymbol{\pi}} \widehat{R}(\boldsymbol{\theta}, \boldsymbol{\pi}) = \min_{\boldsymbol{\theta},\boldsymbol{\pi}} \widehat{E_{X,Y}} E_{\widetilde{S}}[L(f_{\boldsymbol{\theta}}(\boldsymbol{X} \odot \widetilde{S}), Y) + \lambda \parallel \widetilde{S} \parallel_0]$$

$$where\ \widetilde{S} = \{0,1\}^D$$

↓

$$\min_{\boldsymbol{\theta},\boldsymbol{\mu}} \widehat{R}(\boldsymbol{\theta}, \boldsymbol{\mu}) = \min_{\boldsymbol{\theta},\boldsymbol{\mu}} \widehat{E_{X,Y}} E_{Z}[L(f_{\boldsymbol{\theta}}(\boldsymbol{X} \odot Z), Y) + \lambda \parallel Z \parallel_0]$$

$$where\ \boldsymbol{Z}\ is\ a\ random\ vector,\ z_d = max(0,\ min(1,\ \sigma_d \epsilon_d + \mu_d + 0.5))$$

Reparametrization Trick

$$\min_{\boldsymbol{\theta},\boldsymbol{\mu}} \widehat{R}(\boldsymbol{\theta},\boldsymbol{\mu}) = \min_{\boldsymbol{\theta},\boldsymbol{\mu}} \widehat{E_{\mathrm{X,Y}}} E_Z[L(f_{\boldsymbol{\theta}}(\boldsymbol{X} \odot Z), Y) + \lambda \parallel Z \parallel_0]$$

$$where\ \boldsymbol{Z}\ is\ a\ random\ vector,\ z_d = max(0,\ min(1,\ \sigma_d\epsilon_d + \mu_d + 0.5))$$

$$\frac{\partial}{\partial \mu_d} E_Z[L(f_{\boldsymbol{\theta}}(\boldsymbol{X} \odot Z), Y) + \lambda \parallel Z \parallel_0] \approx \frac{1}{K}\sum_{k=1}^{K}[L^{'}(\boldsymbol{z}^k)\frac{\partial z_d^k}{\partial \mu_d}] + \lambda \frac{\partial}{\partial \mu_d}\sum_{d=1}^{D} Pr\{z_d > 0\}$$

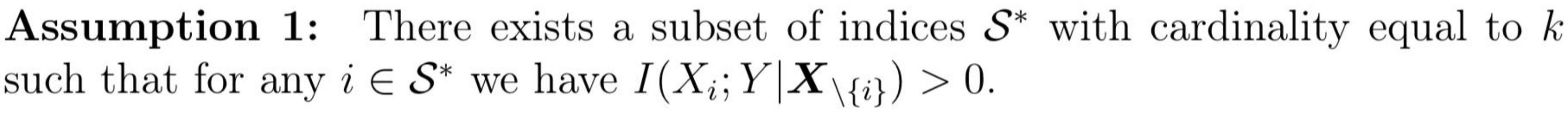
In inference stage, if $\widehat{z_d}$ is greater than a threshold(e.g. 0.5), then select $\widehat{z_d}$ as a relevant feature, otherwise drop it.

Mutual Information Perspective

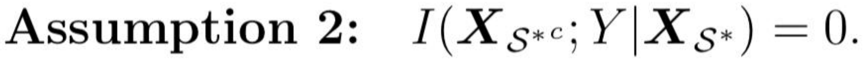
To select a subset of features S of size k that has the highest mutual information with the target variable Y.

$$\max_{S} I(\boldsymbol{X}_S,\ Y)\ \ s.t.\ |S| = k$$

Connection to Mutual Information

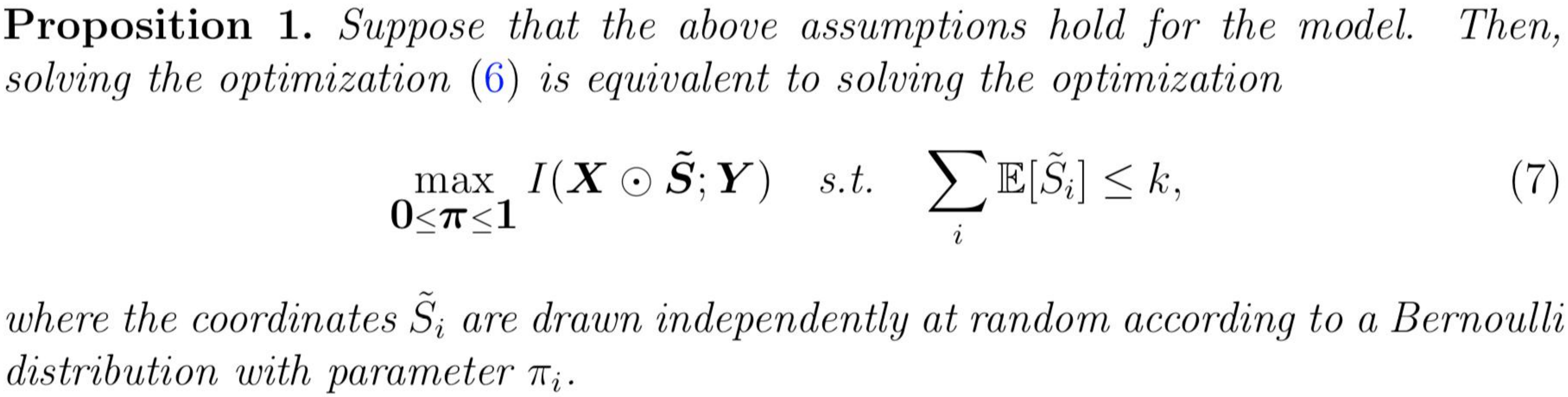


Assumption 1: There exists a subset of indices $\mathcal{S}^*$ with cardinality equal to k such that for any $i \in \mathcal{S}^*$ we have $I(X_i; Y | \boldsymbol{X}_{\setminus\{i\}}) > 0$.



Assumption 2: $I(\boldsymbol{X}_{\mathcal{S}^{*c}}; Y | \boldsymbol{X}_{\mathcal{S}^*}) = 0$.

$$\max_{S} I(\boldsymbol{X}_S,\ Y)\ \ s.t.\ |S| = k \quad (6)$$

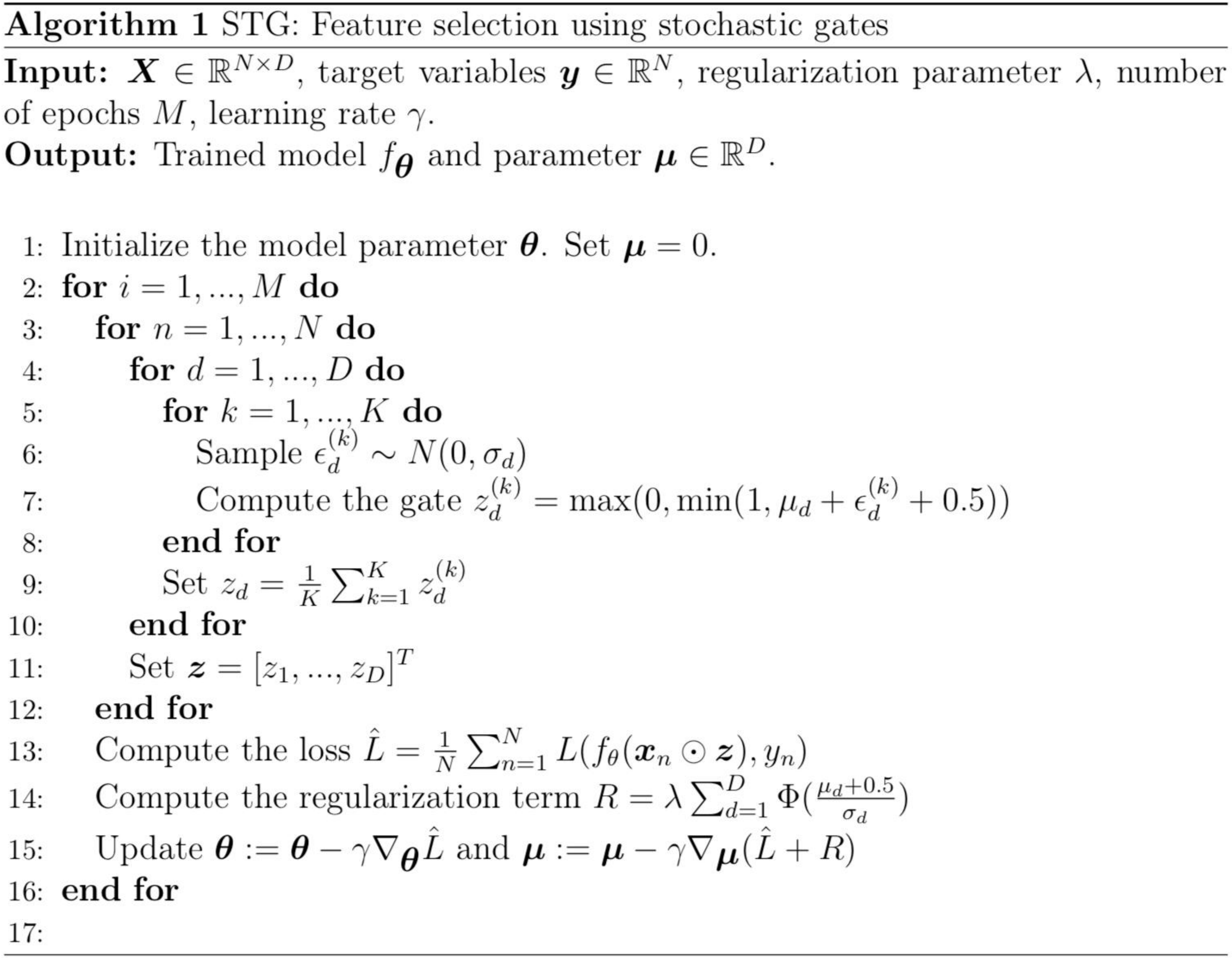


Proposition 1. *Suppose that the above assumptions hold for the model. Then, solving the optimization (6) is equivalent to solving the optimization*

$$\max_{\mathbf{0} \leq \boldsymbol{\pi} \leq \mathbf{1}} I(\boldsymbol{X} \odot \tilde{\boldsymbol{S}}; \boldsymbol{Y}) \quad s.t. \quad \sum_i \mathbb{E}[\tilde{S}_i] \leq k, \tag{7}$$

where the coordinates $\tilde{S}_i$ are drawn independently at random according to a Bernoulli distribution with parameter π_i.

Algorithm



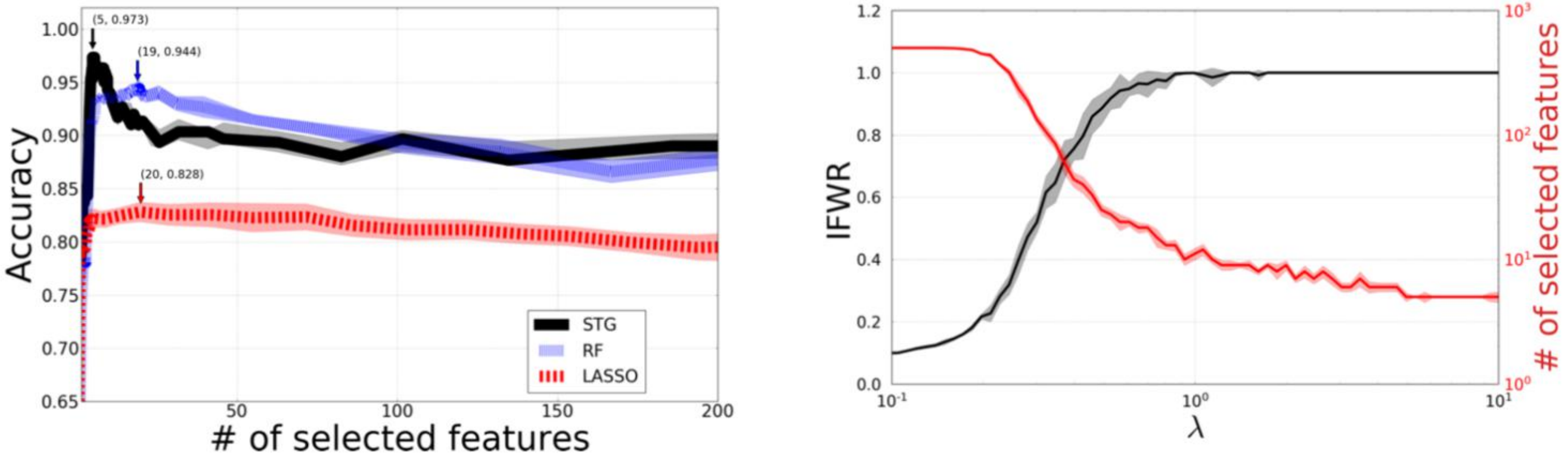
Algorithm 1 STG: Feature selection using stochastic gates

Input: $\boldsymbol{X} \in \mathbb{R}^{N \times D}$, target variables $\boldsymbol{y} \in \mathbb{R}^{N}$, regularization parameter λ, number of epochs M, learning rate γ.

Output: Trained model $f_{\boldsymbol{\theta}}$ and parameter $\boldsymbol{\mu} \in \mathbb{R}^{D}$.

```
Initialize the model parameter $\boldsymbol{\theta}$. Set $\boldsymbol{\mu} = 0$.
for $i = 1, ..., M$ do
    for $n = 1, ..., N$ do
        for $d = 1, ..., D$ do
            for $k = 1, ..., K$ do
                Sample $\epsilon_d^{(k)} \sim N(0, \sigma_d)$
                Compute the gate $z_d^{(k)} = \max(0, \min(1, \mu_d + \epsilon_d^{(k)} + 0.5))$
            end for
            Set $z_d = \frac{1}{K} \sum_{k=1}^{K} z_d^{(k)}$
        end for
        Set $\boldsymbol{z} = [z_1, ..., z_D]^T$
    end for
    Compute the loss $\hat{L} = \frac{1}{N} \sum_{n=1}^{N} L(f_\theta(\boldsymbol{x}_n \odot \boldsymbol{z}), y_n)$
    Compute the regularization term $R = \lambda \sum_{d=1}^{D} \Phi(\frac{\mu_d + 0.5}{\sigma_d})$
    Update $\boldsymbol{\theta} := \boldsymbol{\theta} - \gamma \nabla_{\boldsymbol{\theta}} \hat{L}$ and $\boldsymbol{\mu} := \boldsymbol{\mu} - \gamma \nabla_{\boldsymbol{\mu}} (\hat{L} + R)$
end for
```

Experiments

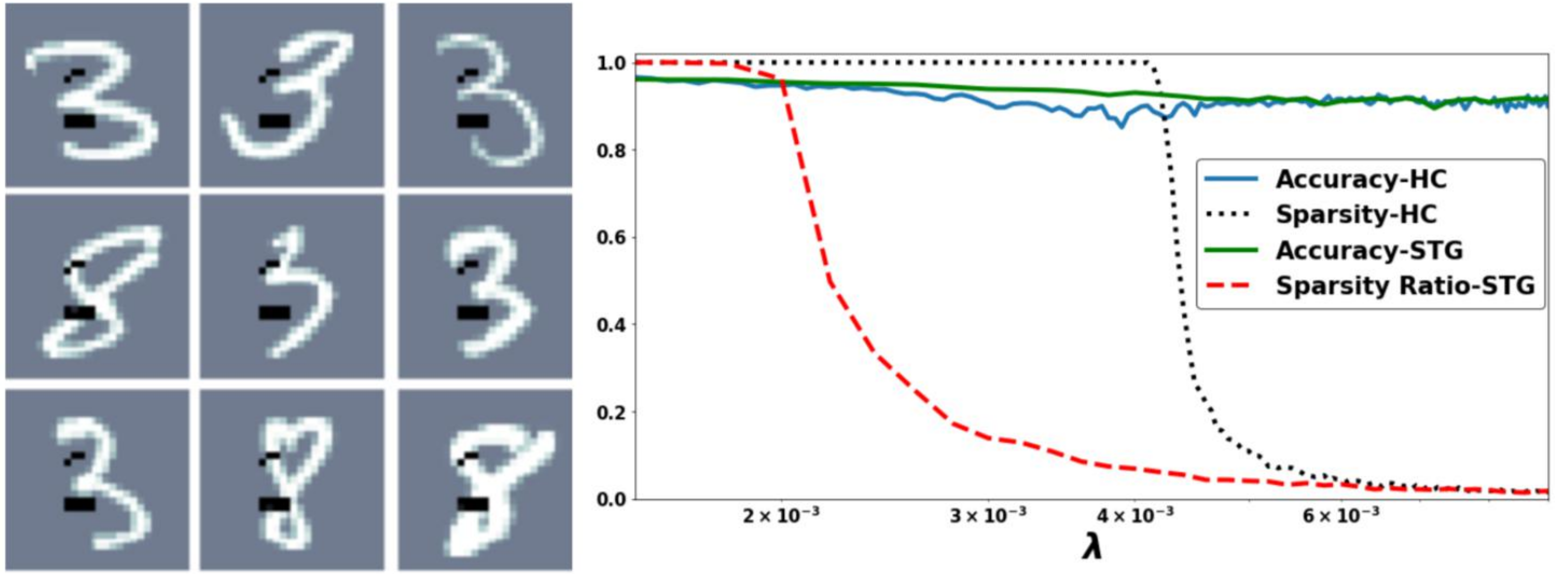


(a)

(b)

Figure 3: (a) Classification accuracy on the MADELON data sets. We evaluate performance using 5-fold cross validation for different number of selected features. In this dataset, only the first 20 coordinates are informative. In that regime the proposed method outperforms RF and LASSO. (b) An empirical evaluation of the effect the regularization parameter λ. The IFWR and the number of selected features are presented on both sides of the y-axis of this plot. For both plots, the mean is presented as a solid/dashed line, while the standard deviation is marked as a shaded color around the mean.

Experiments



(a) (b)

Figure 6: (a) Nine samples from MNIST (white) overlaid with the subset of 13 features (black) selected by STG. Based on only these features, the binary classification accuracy reaches 92.2%. For these nine randomly selected samples, all the 8's have values within the support of the selected features, whereas for the 3's there is no intersection. (b) The comparison of accuracy and sparsity level performance for λ in the range of $[10^{-3}, 10^{-2}]$ between using our proposed method (STG) and its variant using the Hard-Concrete (HC) distribution.

References

- 1. Chen, Jianbo, et al. "Kernel feature selection via conditional covariance minimization." Advances in Neural Information Processing Systems. 2017.
- 2. Yutaro, Yamada, et al. "Feature selection using Stochastic Gates" arXiv preprint arXiv:1810.04247v3 (2019).
- 3. Wasserman, Larry. All of statistics: a concise course in statistical inference. Springer Science & Business Media, 2013.

Q&A