Softplus Regressions and Convex Polytopes

Mingyuan Zhou

IROM Department, McCombs School of Business
The University of Texas at Austin

The 10th ICSA International Conference
Shanghai Jiaotong University
Shanghai, China, December 21, 2016

Binary classification

- Linear classifier:
  - Logistic regression
  - Probit regression
  - Use a single hyperplane to partition the covariate space into two halves
- Nonlinear classifier:
  - Use the kernel trick:
    - Choose a subset of covariate vectors as support vectors
    - Compute a sample's kernel distances to these support vectors
    - Regress the label on the kernel distances
    - Often not scalable
  - Use a deep neural network
    - Transform the covariates with a deep neural network
    - Regress on the transformed covariates
    - Need to tune the network structure

Sum-stack-softplus regression on “XOR”

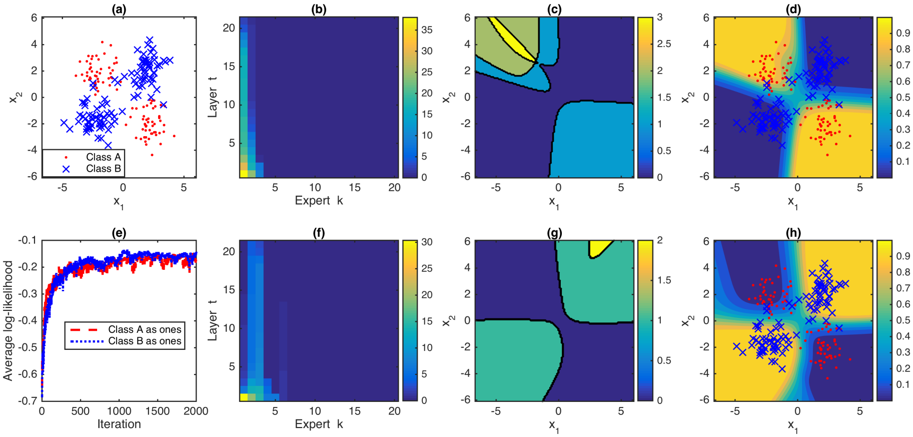


Figure: Visualization of sum-stack-softplus regression, with $K_{\max} = 20$ experts and $T = 20$ criteria for each expert, on classifying the XOR dataset under two opposite labeling settings.

Motivations

- Exploit two distinct types of interactions—noisy-OR and noisy-AND—between hyperplanes to define flexible nonlinear classification decision boundaries directly on the original covariate space.
- Attribute a binary outcome to multiple hidden causes, each of which is associated with an activation probability function produced by a single hyperplane or the collaboration of multiple ones.
- The noisy-OR and/or noisy-AND interactions of hyperplanes make it simple to interpret and quantify how each hyperplane contributes to the final classification decision boundaries.
- Provide probability estimates, automatically learn the complexity of the predictive distribution, and quantify model uncertainties.

Bernoulli-Poisson link

- Bernoulli-Poisson link:

$$y = \delta(m \geq 1), \; m \sim \mathrm{Pois}(\lambda)$$

- The marginalization of the latent count m leads to

$$y \sim \mathrm{Bernoulli}(p), \; p = 1 - e^{-\lambda}$$

  where $\lambda = -\ln(1-p)$ is referred to as the Bernoulli-Poisson rate.

- Since $1/(1+e^{-x}) = 1 - \exp[-\ln(1+e^{x})]$, letting

$$y \sim \mathrm{Bernoulli}\big[\sigma(x)\big], \; \sigma(x) = 1/(1+e^{-x})$$

  is equivalent to letting

$$y \sim \mathrm{Bernoulli}\big(1 - e^{-\varsigma(x)}\big), \; \varsigma(x) = \ln(1+e^{x}).$$

Softplus function

- $\varsigma(x) = \ln(1 + e^x)$ is the softplus function.
- The softplus function is a smoothed version of the rectifier, or rectified linear unit
$$\mathrm{ReLU}(x) = \max(0, x).$$
- The rectifier function is now widely used in deep neural networks, replacing other canonical nonlinear activation functions such as the sigmoid and hyperbolic tangent functions.

A family of softplus functions

- Stack-softplus function:

$$\varsigma(x_1,\ldots,x_t) = \ln\left(1 + e^{x_t}\ln\left\{1 + e^{x_{t-1}}\ln\left[1 + \ldots\ln\left(1 + e^{x_1}\right)\right]\right\}\right)$$

Recursive definition: $\varsigma(x_1,\ldots,x_t) = \ln[1 + e^{x_t}\varsigma(x_1,\ldots,x_{t-1})]$.

- Sum-softplus:

$$\sum\nolimits_{k=1}^{\infty} r_k\,\varsigma(x_k),$$

where r_k are the countably infinite weights of a gamma process.

- Sum-stack-softplus (SS-softplus) function:

$$\sum\nolimits_{k=1}^{\infty} r_k\,\varsigma(x_{k1},\ldots,x_{kt}).$$

- While the softplus function is monotonic, the stack-, sum-, and SS-softplus functions could produce a single peak, a single valley, and multiple change points, respectively, along the real line.

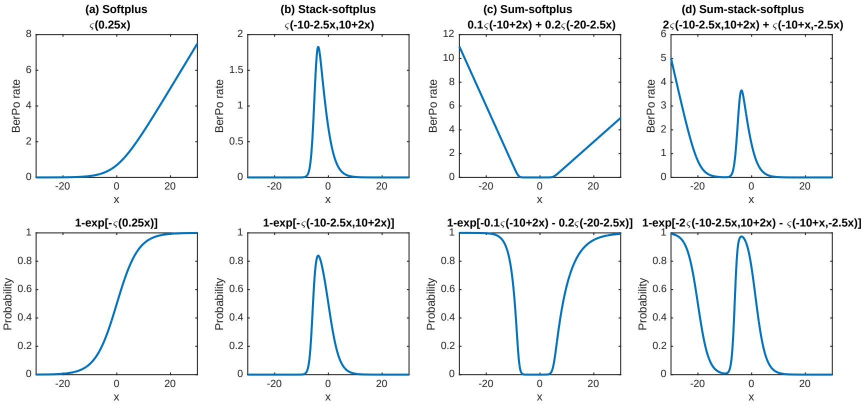


Figure: First row: $\lambda(x)$. Second row: $1 - e^{-\lambda(x)}$.

A family of softplus regressions

- For the ith covariate vector $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{iV})' \in \mathbb{R}^{V+1}$, we model its binary class label $y_i \in \{0, 1\}$ using

$$y_i \,|\, \boldsymbol{x}_i \sim \text{Bernoulli}(1 - e^{-\lambda(\boldsymbol{x}_i)})$$

- $\lambda(\boldsymbol{x}_i)$ is a nonnegative deterministic function of $\boldsymbol{x}_i$ that may contain countably infinite parameters drawn from a completely random measure.

Definition (Sum-softplus regression)

Given a gamma process draw $G = \sum_{k=1}^{\infty} r_k \delta_{\boldsymbol{\beta}_k}$, sum-softplus regression parameterizes $\lambda(\boldsymbol{x}_i)$ using a sum-softplus function as

$$\lambda(\boldsymbol{x}_i) = \sum_{k=1}^{\infty} r_k \,\varsigma(\boldsymbol{x}_i' \boldsymbol{\beta}_k) = \sum_{k=1}^{\infty} r_k \ln(1 + e^{\boldsymbol{x}_i' \boldsymbol{\beta}_k}).$$

- Sum-softplus regression is equivalent to a noisy-OR binary regression model

$$y_i \sim \text{Bernoulli}\left[1 - \prod_{k=1}^{\infty}(1 - p_{ik})\right], \quad p_{ik} = 1 - \left(\frac{1}{1 + e^{\boldsymbol{x}_i' \boldsymbol{\beta}_k}}\right)^{r_k}.$$

- It can be constructed using the hierarchical model

$$y_i = \delta(m_i \geq 1),\ m_i \sim \text{Pois}(\theta_i),\ \theta_i = \sum_{k=1}^{\infty} \theta_{ik},\ \theta_{ik} \sim \text{Gamma}\big(r_k, e^{\boldsymbol{x}_i' \boldsymbol{\beta}_k}\big).$$

- Sum-softplus regression can also be constructed with

$$y_i = \delta(m_i \geq 1), \quad m_i = \sum_{k=1}^{\infty} m_{ik}, \ m_{ik} \sim \mathrm{NB}\left[r_k, 1/(1+e^{-\boldsymbol{x}_i'\boldsymbol{\beta}_k})\right],$$

Proposition

The infinite product

$$e^{-\sum_{k=1}^{\infty} r_k \, \varsigma(\boldsymbol{x}_i'\boldsymbol{\beta}_k)} = \prod_{k=1}^{\infty} \left(1 + e^{\boldsymbol{x}_i'\boldsymbol{\beta}_k}\right)^{-r_k}$$

in sum-softplus regression is smaller than one and has a finite expectation that is greater than zero.

Geometric constraint of sum-softplus regression

- uses the **noisy-OR** hyperplane interactions
  to define a **convex-polytope**-bounded confined space
  to enclose **negative** examples (i.e., data samples with $y_i = 0$)

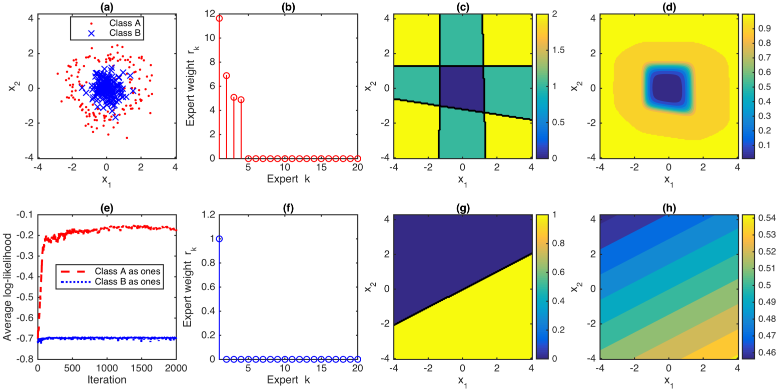


Figure:

First row: Red and Blue points are labeled as “1” and “0,” respectively.

Second row: Blue and Red points are relabeled as “1” and “0,” respectively.

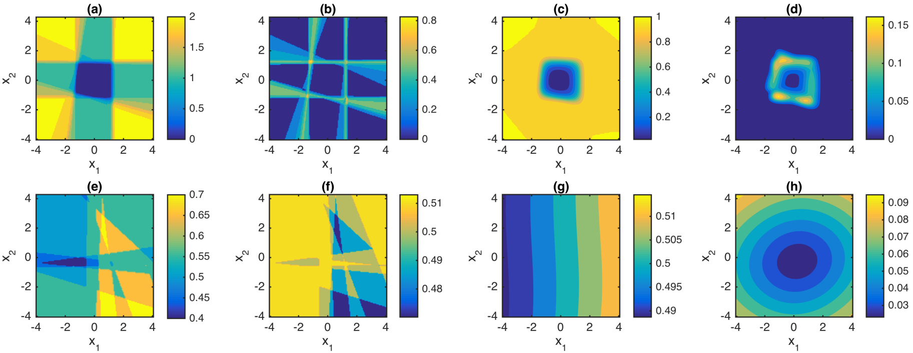


Figure: Visualization of the posteriors of sum-softplus regression based on 20 MCMC samples, collected once per every 50 iterations during the last 1000 MCMC iterations. (a) and (b) show the contour maps of the posterior means and standard deviations, respectively, of the number of "Yes" votes, and (c) and (d) show the contour maps of the posterior means and standard deviations, respectively, of predicted class probabilities. (e)-(h) are analogous plots to (a)-(d), with the data points in Classes A and B relabeled as "0" and "1," respectively.

Definition (Stack-softplus regression)

With weight $r \in \mathbb{R}_+$ and T regression coefficient vectors $\boldsymbol{\beta}^{(2:T+1)} := (\boldsymbol{\beta}^{(2)}, \dots, \boldsymbol{\beta}^{(T+1)}) \in \mathbb{R}^{(V+1)\times T}$, stack-softplus regression with T layers parameterizes $\lambda(\boldsymbol{x}_i)$ using a stack-softplus function as

$$\begin{aligned}\lambda(\boldsymbol{x}_i) &= r\,\varsigma\big(\boldsymbol{x}_i'\boldsymbol{\beta}^{(2)}, \dots, \boldsymbol{x}_i'\boldsymbol{\beta}^{(T+1)}\big)\\ &= r \ln\left(1 + e^{\boldsymbol{x}_i'\boldsymbol{\beta}^{(T+1)}} \ln\left\{1 + e^{\boldsymbol{x}_i'\boldsymbol{\beta}^{(T)}} \ln\left[1 + \dots \ln\left(1 + e^{\boldsymbol{x}_i'\boldsymbol{\beta}^{(2)}}\right)\right]\right\}\right).\end{aligned}$$

- Stack-softplus regression is equivalent to a noisy-AND regression model

$$\begin{aligned} y_i &\sim \mathrm{Bernoulli}(p_i),\\ p_i &= 1 - \left(1 + e^{\boldsymbol{x}_i'\boldsymbol{\beta}^{(T+1)}} \ln\left\{1 + e^{\boldsymbol{x}_i'\boldsymbol{\beta}^{(T)}} \ln\left[1 + \dots \ln\left(1 + e^{\boldsymbol{x}_i'\boldsymbol{\beta}^{(2)}}\right)\right]\right\}\right)^{-r}.\end{aligned}$$

- Stack-softplus regression can be constructed using the hierarchical model that stacks T gamma distributions, whose scales are differently parameterized by the covariates, as

$$\theta_i^{(T)} \sim \text{Gamma}\left(r, e^{\boldsymbol{x}_i' \boldsymbol{\beta}^{(T+1)}}\right),$$

$$\cdots$$

$$\theta_i^{(t)} \sim \text{Gamma}\left(\theta_i^{(t+1)}, e^{\boldsymbol{x}_i' \boldsymbol{\beta}^{(t+1)}}\right),$$

$$\cdots$$

$$y_i = \delta(m_i \geq 1),\ m_i \sim \text{Pois}(\theta_i^{(1)}),\ \theta_i^{(1)} \sim \text{Gamma}\left(\theta_i^{(2)}, e^{\boldsymbol{x}_i' \boldsymbol{\beta}^{(2)}}\right).$$

Geometric constraint of stack-softplus regression

- uses the **noisy-AND** hyperplane interactions
  to define a **convex-polytope**-like confined space
  to enclose **positive** examples (i.e., data samples with $y_i = 1$)

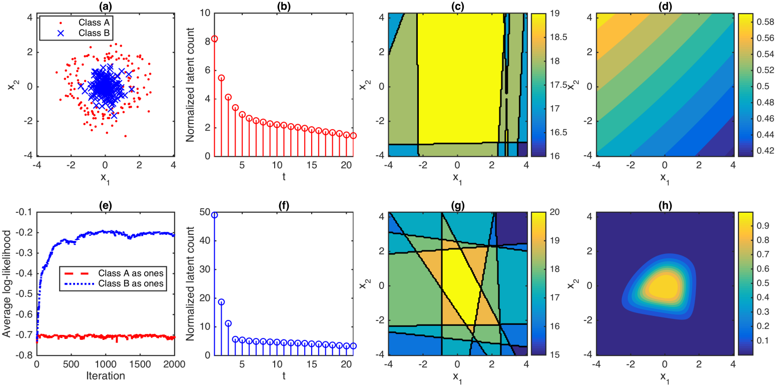


Figure:
First row: Red and Blue points are labeled as “1” and “0,” respectively.
Second row: Blue and Red points are relabeled as “1” and “0,” respectively.

Definition (Sum-stack-softplus regression)

Given a gamma process draw $G = \sum_{k=1}^{\infty} r_k \delta_{\boldsymbol{\beta}_k^{(2:T+1)}}$, with each $\boldsymbol{\beta}_k^{(t)} \in \mathbb{R}^{V+1}$, sum-stack-softplus (SS-softplus) regression with $T \in \{1, 2, \ldots\}$ layers parameterizes $\lambda(\boldsymbol{x}_i)$ using a SS-softplus function as

$$\lambda(\boldsymbol{x}_i) = \sum_{k=1}^{\infty} r_k \, \varsigma\big(\boldsymbol{x}_i' \boldsymbol{\beta}_k^{(2)}, \ldots, \boldsymbol{x}_i' \boldsymbol{\beta}_k^{(T+1)}\big)$$

$$= \sum_{k=1}^{\infty} r_k \ln \left(1 + e^{\boldsymbol{x}_i' \boldsymbol{\beta}_k^{(T+1)}} \ln \left\{1 + e^{\boldsymbol{x}_i' \boldsymbol{\beta}_k^{(T)}} \ln \left[1 + \ldots \ln \left(1 + e^{\boldsymbol{x}_i' \boldsymbol{\beta}_k^{(2)}}\right)\right]\right\}\right).$$

- SS-softplus regression is equivalent to the following noisy-OR-AND regression model

$$\begin{aligned} y_i &\sim \text{Bernoulli}\left[1 - \prod_{k=1}^{\infty} (1 - p_{ik})\right], \\ p_{ik} &= 1 - \left(1 + e^{\boldsymbol{x}_i' \boldsymbol{\beta}_k^{(T+1)}} \ln\left\{1 + e^{\boldsymbol{x}_i' \boldsymbol{\beta}_k^{(T)}} \ln\left[1 + \ldots \ln\left(1 + e^{\boldsymbol{x}_i' \boldsymbol{\beta}_k^{(2)}}\right)\right]\right\}\right)^{-r_k} \end{aligned}$$

- Sum-stack-softplus regression can be constructed by convolving countably infinite stacked gamma distributions that have covariate-dependent scale parameters as

$$\begin{gathered}
\theta_{ik}^{(T)} \sim \text{Gamma}\left(r_k, e^{\boldsymbol{x}_i'\boldsymbol{\beta}_k^{(T+1)}}\right), \\
\cdots \\
\theta_{ik}^{(t)} \sim \text{Gamma}\left(\theta_{ik}^{(t+1)}, e^{\boldsymbol{x}_i'\boldsymbol{\beta}_k^{(t+1)}}\right), \\
\cdots \\
\theta_{ik}^{(1)} \sim \text{Gamma}\left(\theta_{ik}^{(2)}, e^{\boldsymbol{x}_j'\boldsymbol{\beta}_k^{(2)}}\right), \\
y_i = \delta(m_i \geq 1),\ m_i = \textstyle\sum_{k=1}^{\infty} m_{ik}^{(1)},\ m_{ik}^{(1)} \sim \text{Pois}(\theta_{ik}^{(1)}).
\end{gathered}$$

Proposition

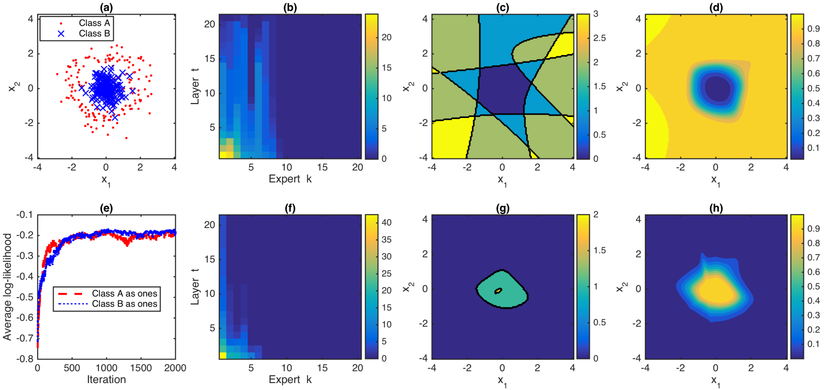
The infinite product in sum-stack-softplus regression as

$$e^{-\sum_{k=1}^{\infty} r_k\, \varsigma\left(\boldsymbol{x}_i'\boldsymbol{\beta}_k^{(2:T+1)}\right)} = \prod_{k=1}^{\infty}\left(1+e^{\boldsymbol{x}_i'\boldsymbol{\beta}_k^{(T+1)}}\ln\left\{1+e^{\boldsymbol{x}_i'\boldsymbol{\beta}_k^{(T)}}\ln\left[1+\ldots\ln\left(1+e^{\boldsymbol{x}_i'\boldsymbol{\beta}_k^{(2)}}\right)\right]\right\}\right)^{-r_k}$$

is smaller than one and has a finite expectation that is greater than zero.

Geometric constraint of sum-stack-softplus regression

- uses the **noisy-OR of noisy-AND** hyperplane interactions
  to define a **union of convex-polytope**-like confined space
  to enclose **positive** examples (i.e., data samples with $y_i = 1$)



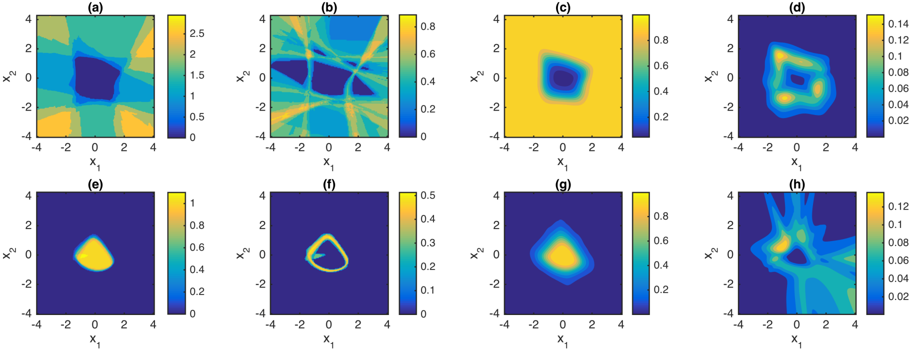


Figure: Analogous figure for sum-stack-softplus regression to that for sum-softplus regression, with the following differences: (a) and (b) show the contour maps of the posterior means and standard deviations, respectively, of the number of "Yes" votes, and (c) and (d) show the contour maps of the posterior means and standard deviations, respectively, of predicted class probabilities. (e)-(h) are analogous plots to (a)-(d), with the data points in Classes *A* and *B* relabeled as "0" and "1," respectively.

Sum-stack-softplus regression on “banana”

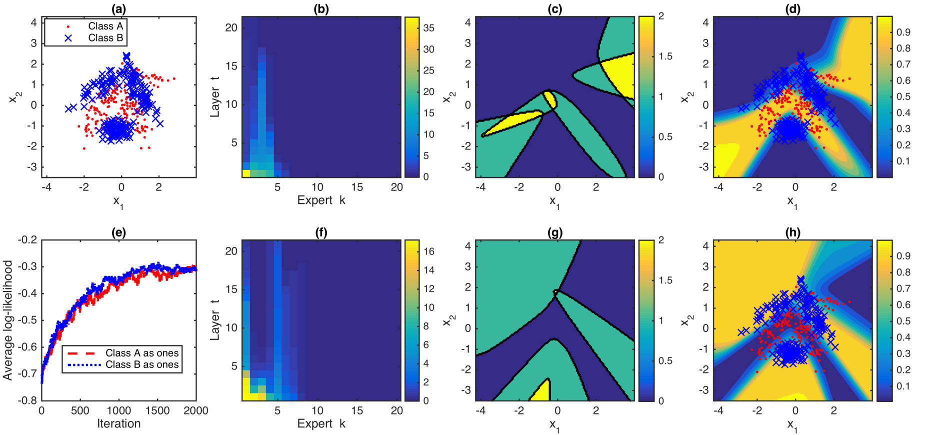


Figure: Visualization of sum-stack-softplus regression, with $K_{\max} = 20$ experts and $T = 20$ criteria for each expert, on classifying the banana dataset under two opposite labeling settings.

Sum-stack-softplus regression on "double moons"

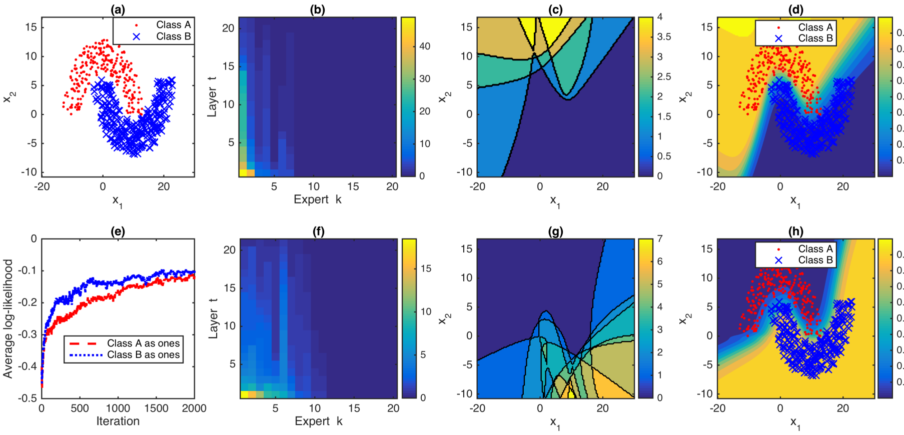


Figure: Visualization of sum-stack-softplus regression, with $K_{\max} = 20$ experts and $T = 20$ criteria for each expert, on classifying the double moons dataset under two opposite labeling settings.

Hierarchical model

- Truncated sum-stack-softplus (SS-softplus) regression:

$$\theta_{ik}^{(T)} \sim \text{Gamma}\left(r_k, e^{\boldsymbol{x}_i' \boldsymbol{\beta}_k^{(T+1)}}\right),$$
$$\cdots$$
$$\theta_{ik}^{(t)} \sim \text{Gamma}\left(\theta_{ik}^{(t+1)}, e^{\boldsymbol{x}_i' \boldsymbol{\beta}_k^{(t+1)}}\right),$$
$$\cdots$$
$$\theta_{ik}^{(1)} \sim \text{Gamma}\left(\theta_{ik}^{(2)}, e^{\boldsymbol{x}_j' \boldsymbol{\beta}_k^{(2)}}\right),$$
$$y_i = \delta(m_i \geq 1),\ m_i = \textstyle\sum_{k=1}^{\infty} m_{ik}^{(1)},\ m_{ik}^{(1)} \sim \text{Pois}(\theta_{ik}^{(1)}).$$

- We complete the model by letting

$$r_k \sim \text{Gamma}(\gamma_0/K, 1/c_0),\ \gamma_0 \sim \text{Gamma}(a_0, 1/b_0),\ c_0 \sim \text{Gamma}(e_0, 1/f_0),$$
$$\boldsymbol{\beta}_k^{(t)} \sim \prod_{v=0}^{V} \mathcal{N}(0, \alpha_{vtk}^{-1}),\ \alpha_{vtk} \sim \text{Gamma}(a_t, 1/b_t),$$

where $t \in \{2, \ldots, T+1\}$

Upward-downward Gibbs sampling

- Closed-form Gibbs sampling update equations via data augmentation and marginalization
- Credit assignment and information propagation via latent counts, which are linked to regression coefficients via negative binomial regressions.

Theorem

One may find latent counts $m_{ik}^{(t)}$ that are connected to the regression coefficient vectors under negative binomial regression as

$$m_{ik}^{(t)} \sim NB(\theta_{ik}^{(t+1)},\ \ 1 - e^{-q_{ik}^{(t+1)}}) = NB\left(\theta_{ik}^{(t+1)}, \frac{1}{1 + e^{-\boldsymbol{x}_i' \boldsymbol{\beta}_k^{(t+1)} - \ln(q_{ik}^{(t)})}}\right)$$

Experiments on benchmark datasets

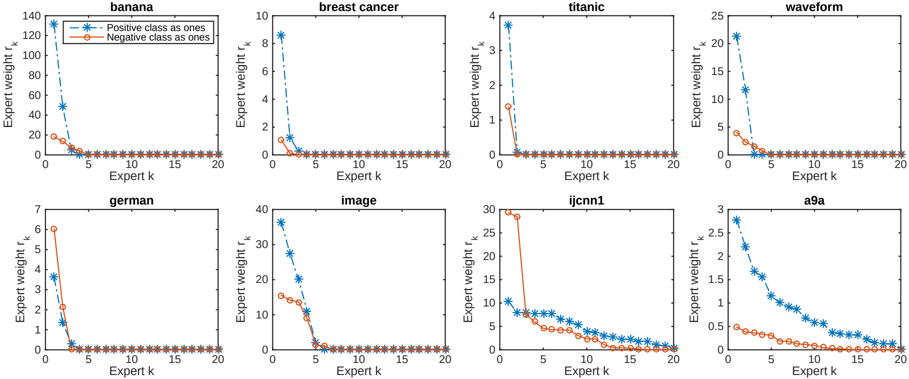


Figure: The inferred weights of the $K_{\max} = 20$ experts with $T = 5$ layers of sum-stack-softplus regression, ordered from left to right according to their weights, on eight different datasets, based on the maximum likelihood sample of a single random trial.

Table: Comparison of classification errors of logistic regression (LR), RBF kernel support vector machine (SVM), relevance vector machine (RVM), adaptive multi-hyperplane machine (AMM), convex polytope machine (CPM), softplus regression, sum-softplus (sum-ς) regression with $K_{\max} = 20$, stack-softplus (stack-ς) regression with $T = 5$, SS-softplus (SS-ς) regression with $K_{\max} = 20$ and $T = 3$, and SS-ς regression with $K_{\max} = 20$ and $T = 5$. Displayed in each column of the last row is the average of the classification errors of an algorithm normalized by those of kernel SVM.

Dataset	LR	SVM	RVM	AMM	CPM	softplus	sum-ς	stack-ς (T=5)	SS-ς (T=3)	SS-ς (T=5)
banana	47.76 ±4.38	**10.85** ±0.57	11.08 ±0.69	18.76 ±4.09	21.39 ±1.72	47.87 ±4.36	30.78 ±8.68	33.21 ±5.76	12.54 ±1.18	11.89 ±0.61
breast cancer	28.05 ±3.68	28.44 ±4.52	31.56 ±4.66	31.82 ±4.47	32.08 ±4.29	28.70 ±4.76	30.13 ±4.23	**27.92** ±3.31	30.39 ±4.94	28.83 ±3.40
titanic	22.67 ±0.98	22.33 ±0.63	23.20 ±1.08	28.85 ±8.56	22.37 ±0.45	22.53 ±0.43	22.48 ±0.25	22.71 ±0.70	22.42 ±0.45	**22.29** ±0.80
waveform	13.33 ±0.59	**10.73** ±0.86	11.16 ±0.72	11.81 ±1.13	12.76 ±1.17	13.62 ±0.71	11.51 ±0.65	12.25 ±0.69	11.34 ±0.70	11.69 ±0.69
german	23.63 ±1.70	23.30 ±2.51	23.67 ±2.28	25.13 ±3.73	25.03 ±2.49	24.07 ±2.11	23.60 ±2.39	**22.97** ±2.22	23.30 ±2.20	24.23 ±2.46
image	17.53 ±1.05	2.84 ±0.52	3.82 ±0.59	3.82 ±0.87	3.25 ±0.41	17.55 ±0.75	3.50 ±0.73	7.97 ±0.52	**2.59** ±0.47	2.73 ±0.53
Mean of SVM normalized errors	2.472	1	1.095	1.277	1.251	2.485	1.370	1.665	1.033	1.033

Table: Comparison of the number of experts (times the number of hyperplanes per expert), where an expert contains T hyperplanes for both stack- and SS-softplus regressions and contains a single hyperplane/support vector for all the others. The computational complexity for out-of-sample prediction is about linear in the number of hyperplanes/support vectors. Displayed in each column of the last row is the average of the number of experts (times the number of hyperplanes per expert) of an algorithm normalized by those of RBF kernel SVM.

Dataset	LR	SVM	RVM	AMM	CPM	softplus	sum-ς	stack-ς (T=5)	SS-ς (T=3)	SS-ς (T=5)
banana	1	129.20 ±32.76	22.30 ±26.02	9.50 ±2.80	14.60 ±7.49	2	3.70 ±0.95	2 (×5)	6.80 (×3) ±0.79 (×3)	7.60 (×5) ±1.17 (×5)
breast cancer	1	115.10 ±11.16	24.80 ±28.32	13.40 ±0.84	12.00 ±8.43	2	3.10 ±0.74	2 (×5)	5.70 (×3) ±1.70 (×3)	6.40 (×5) ±1.43 (×5)
titanic	1	83.40 ±13.28	5.10 ±3.03	14.90 ±3.14	5.20 ±2.53	2	2.30 ±0.48	2 (×5)	3.80 (×3) ±0.92 (×3)	4.00 (×5) ±0.94 (×5)
waveform	1	147.00 ±38.49	21.10 ±10.98	9.50 ±1.18	6.40 ±2.27	2	4.40 ±0.84	2 (×5)	7.00 (×3) ±2.21 (×3)	8.90 (×5) ±2.33 (×5)
german	1	423.60 ±55.02	11.00 ±3.20	18.80 ±1.81	8.80 ±7.79	2	6.70 ±0.95	2 (×5)	11.10 (×3) ±2.64 (×3)	14.70 (×5) ±1.77 (×5)
image	1	211.60 ±47.51	35.80 ±9.19	10.50 ±1.08	23.00 ±6.75	2	11.20 ±1.32	2 (×5)	14.60 (×3) ±2.07 (×3)	17.60 (×5) ±1.90 (×5)
Mean of SVM normalized K	0.007	1	0.131	0.088	0.075	0.014	0.030	0.014 (×5)	0.048 (×3)	0.057 (×5)

Conclusions

- We propose sum-, stack-, and sum-stack-softplus regressions that combine multiple hyperplanes, respectively,
  - via the noisy-OR interaction to construct a convex-polytope-bounded confined space to enclose the negative class,
  - via the noisy-AND interaction to construct a convex-polytope-bounded confined space to enclose the negative class,
  - and via the noisy-OR-AND interaction to construct a union of convex-polytope-like confined spaces to enclose the positive class.
- Sum-stack-softplus regression constructs a highly flexible nonparametric Bayesian predictive distribution by mixing the convolved and stacked covariate-dependent gamma distributions with the Bernoulli-Poisson distribution.

- The predictive distribution is deconvolved and demixed by inferring the parameters of the underlying nonparametric Bayesian hierarchical model using a series of data augmentation and marginalization techniques.
- Example results demonstrate that the proposed softplus regressions
  - can achieve classification accuracies comparable to those of kernel support vector machine,
  - but consume significant less computation for out-of-sample predictions,
  - provide probability estimates, quantify uncertainties,
  - and place interpretable geometric constraints on its classification decision boundaries directly in the original covariate space.