Recap for Final

Sreyas Mohan

CDS at NYU

8 May 2019

Learning Theory Framework

1. Input Space $\mathcal{X}$, output space $\mathcal{Y}$ and Action Space $\mathcal{A}$
2. Prediction function $f : \mathcal{X} \to \mathcal{A}$
3. Loss Function $\ell : \mathcal{A} \times \mathcal{Y} \to \mathcal{R}$
4. Risk $R(f) = E(\ell(f(x), y)$
5. Bayes prediction function $f^* = \text{argmin}_f R(f)$
6. Bayes Risk $R(f^*)$
7. Empirical Risk,$\hat{R}(f)$ - sample approximation for $R(f)$. $\hat{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n}(f(x_i), y_i)$
8. Empirical Risk Minimization $\hat{f} = \text{argmin}_f \hat{R}(f)$
9. Constrained Empirical Risk Minization

Learning Theory Framework

$$f^* = \arg\min_f \ell(f(X), Y)$$

$$f_{\mathcal{F}} = \arg\min_{f \in \mathcal{F}} \ell(f(X), Y))$$

$$\hat{f}_n = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i)$$

- Approximation Error (of $\mathcal{F}$) = $R(f_{\mathcal{F}}) - R(f^*)$
- Estimation error (of $\hat{f}_n$ in F) = $R(\hat{f}_n) - R(f_{\mathcal{F}})$

Approximation Error of Regression Stumps

Suppose $\mathcal{X} = \mathcal{U}[-1, 1]$. $y|x \sim \mathcal{N}(x, \sigma^2)$. We are minimizing square loss. Consider a hypothesis class $\mathcal{H}$ made of regression stumps. What is the $f \in \mathcal{H}$ that minimizes the approximation error of $\mathcal{H}$?

Solution

Note that $f^*(x) = x$ and $R(f^*) = \sigma^2$.
Also $R(f) = E((f(x) - y)^2) = E[(f(x) - E[y|x])^2] + E[(y - E[y|x])^2]$
We have to minimize $E[(f(x) - E[y|x])^2] = E[(f(x) - x)^2]$
The function we are looking for is

$$f(x) = \begin{cases} -0.5 & -1 \leq x \leq 0 \\ +0.5 & 0 < x \leq 1 \end{cases}$$

Lasso and Ridge Regression

- Linear Regression Closed form Expression ($w = X^T X)^{-1} X^T y$
- Ridge Regression Solution
- Lasso Regression Solution Methods - co-ordinate descent.
- Correlated Features for Lasso and Ridge

Optimization Methods

- (True/False) Gradient Descent or SGD is only defined for convex loss functions.
- (True/False) Negative Subgradient direction is a descent direction
- (True/False) Negative Stochastic Gradient direction is a descent direction

Optimization Methods

- False
- False, but takes you closed to minimizer
- False

SVM, RBF Kernel and Neural Network

Suppose we have fit a kernelized SVM to some training data $(x_1, y_1), \ldots, (x_n, y_n) \in \times \{-1, 1\}$, and we end up with a score function of the form

$$f(x) = \sum_{i=1}^{n} \alpha_i k(x, x_i),$$

where $k(x, x') = \varphi(x - x')$, for some function $\varphi : \mathbb{R} \to \mathbb{R}$. Write $f : \mathbb{R} \to \mathbb{R}$ as a multilayer perceptron by specifying how to set m and a_i, w_i, b_i for $i = 1, \ldots, m$, and the activation function $\sigma : \mathbb{R} \to \mathbb{R}$ in the following expression:

$$f(x) = \sum_{i=1}^{m} a_i \sigma(w_i x + b_i)$$

Solution

Simply take $m = n$. Then $a_i = \alpha_i$, $w_i = 1$, and $b_i = -x_i$. Finally take $\sigma = \varphi$.

Question continued

Suppose you knew which all data points were support vectors. How would you use this information to reduce the size of the network?

Solution

Remove all the nodes corresponding to $a_i = 0$ or non-support vectors.

Maximum Likelihood

Suppose we have samples $x_1 \leq x_2 \leq \cdots \leq x_n$ drawn from $\mathcal{U}[a, b]$ Find Maximum Likelihood estimate of a and b.

ML Solution

- The likelihood is:

$$L(a, b) = \Pi_{i=1}^{n} \left(\frac{1}{b-a} 1_{[a,b]}(x_i) \right)$$

- Let $x_{(1)}, \ldots, x_{(n)}$ be the order statistics.
- The likelihood is greater than zero if and only $a < x_{(1)}$ and $b > x_{(n)}$.
- When $a < x_{(1)}$ and $b > x_{(n)}$, the likelihood is a monotonically decreasing function of $(b - a)$.
- And the smallest $(b - a)$ will be attained when $b = x_{(n)}$ and $a = x_{(1)}$.
- Therefore, $b = x_{(n)}$ and $a = x_{(1)}$ give us the MLE.

Bayesian

Bayesian Bernoulli Model

Suppose we have a coin with unknown probability of heads $\theta \in (0, 1)$. We flip the coin n times and get a sequence of coin flips with n_h heads and n_t tails.

Recall the following: A Beta(α, β) distribution, for shape parameters $\alpha, \beta > 0$, is a distribution supported on the interval $(0, 1)$ with PDF given by

$$f(x; \alpha, \beta) \propto x^{\alpha-1}(1-x)^{\beta-1}.$$

The mean of a Beta(α, β) distribution is $\frac{\alpha}{\alpha+\beta}$. The mode is $\frac{\alpha-1}{\alpha+\beta-2}$ assuming $\alpha, \beta \geq 1$ and $\alpha + \beta > 2$. If $\alpha = \beta = 1$, then every value in $(0, 1)$ is a mode.

Question Continued

Which ONE of the following prior distributions on θ corresponds to a strong belief that the coin is approximately fair (i.e. has an equal probability of heads and tails)?

1. Beta$(50, 50)$
2. Beta$(0.1, 0.1)$
3. Beta$(1, 100)$

Question Continued

1. Give an expression for the likelihood function $L_{\mathcal{D}}(\theta)$ for this sequence of flips.
2. Suppose we have a Beta(α, β) prior on θ, for some $\alpha, \beta > 0$. Derive the posterior distribution on θ and, if it is a Beta distribution, give its parameters.
3. If your posterior distribution on θ is Beta$(3, 6)$, what is your MAP estimate of θ?

- $L_{\mathcal{D}}(\theta) = \theta^{n_h}(1-\theta)^{n_t}$
-

$$\begin{aligned} p(\theta|\mathcal{D}) &\propto p(\theta)L_(\theta) \\ &\propto \theta^{\alpha-1}(1-\theta)^{\beta-1}\theta^{n_h}(1-\theta)^{n_t} \\ &\propto \theta^{n_h+\alpha-1}(1-\theta)^{n_t+\beta-1} \end{aligned}$$

  Thus the posterior distribution is Beta$(\alpha + n_h, \beta + n_t)$.
- Based on information box above, the mode of the beta distribution is $\frac{\alpha-1}{\alpha+\beta-2}$ for $\alpha, \beta > 1$. So the MAP estimate is $\frac{2}{7}$.

Boostrap

- What is the probability of not picking one datapoint while creating a bootstrap sample?
- Suppose the dataset is fairly large. In an expected sense, what fraction of our bootstrap sample will be unique?

Bootstrap Solutions

1. $(1 - \frac{1}{n})^n$
2. As $n \to \infty$, $(1 - \frac{1}{n})^n \to \frac{1}{e}$. So $1 - \frac{1}{e}$ unique samples.

Random Forest and Boosting

Indicate whether each of the statements (about random forests and gradient boosting) is true or false.

1. True or False: If your gradient boosting model is overfitting, taking additional steps is likely to help.
2. True or False: In gradient boosting, if you reduce your step size, you should expect to need fewer rounds of boosting (i.e. fewer steps) to achieve the same training set loss.
3. True or False: Fitting a random forest model is extremely easy to parallelize.
4. True or False: Fitting a gradient boosting model is extremely easy to parallelize, for any base regression algorithm.
5. True or False: Suppose we apply gradient boosting with absolute loss to a regression problem. If we use linear ridge regression as our base regression algorithm, the final prediction function from gradient boosting always will be an affine function of the input.

Solution

False, False, True, False, True

Hypothesis space of GBM and RF

Let $\mathcal{H}_B$ represent a base hypothesis class of (small) regression trees. Let $\mathcal{H}_R = \{g | g = \sum_{i=1}^{T} \frac{1}{T} f_i, f_i \in \mathcal{H}_B\}$ represent the hypothesis space of prediction functions in a random forest with T trees where each tree is picked from $\mathcal{H}_B$. Let $\mathcal{H}_G = \{g | g = \sum_{i=1}^{T} \nu_i f_i, f_i \in \mathcal{H}_B, n_i \in \mathbb{R}\}$ represent the hypothesis space of prediction functions in a gradient boosting with T trees.

True or False:

1. If $f_i \in \mathcal{H}_R$ then $\alpha f_i \in \mathcal{H}_R$ for all $\alpha \in \mathbb{R}$
2. If $f_i \in \mathcal{H}_G$ then $\alpha f_i \in \mathcal{H}_G$ for all $\alpha \in \mathbb{R}$
3. If $f_i \in \mathcal{H}_G$ then $f_i \in \mathcal{H}_R$
4. If $f_i \in \mathcal{H}_R$ then $f_i \in \mathcal{H}_G$

Solutions

True, True, True, True

Neural Networks

1. True or False: Consider a hypothesis space $\mathcal{H}$ of prediction functions $f : \mathbb{R}^d \to \mathbb{R}$ given by a multilayer perceptron (MLP) with 3 hidden layers, each consisting of m nodes, for which the activation function is $\sigma(x) = cx$, for some fixed $c \in \mathbb{R}$. Then this hypothesis space is strictly larger than the set of all affine functions mapping $\mathbb{R}^d$ to $\mathbb{R}$.
2. True or False: Let $g : [0,1]^d \to \mathbb{R}$ be any continuous function on the compact set $[0,1]^d$. Then for any > 0, there exists $m \in \{1, 2, 3, \ldots\}$, $a = (a_1, \ldots, a_m) \in^m, b = (b_1, \ldots, b_m) \in^m$, and

$$W = \begin{pmatrix} - & w_1^T & - \\ \vdots & \vdots & \vdots \\ - & w_m^T & - \end{pmatrix} \in \mathbb{R}^{m \times d}$$ for which the function $f : [0,1]^d \to \mathbb{R}$ given by

$$f(x) = \sum_{i=1}^{m} a_i \max(0, w_i^T x + b_i)$$

satisfies $|f(x) - g(x)| < \epsilon$ for all $x \in [0,1]^d$.

Solutions

False, True

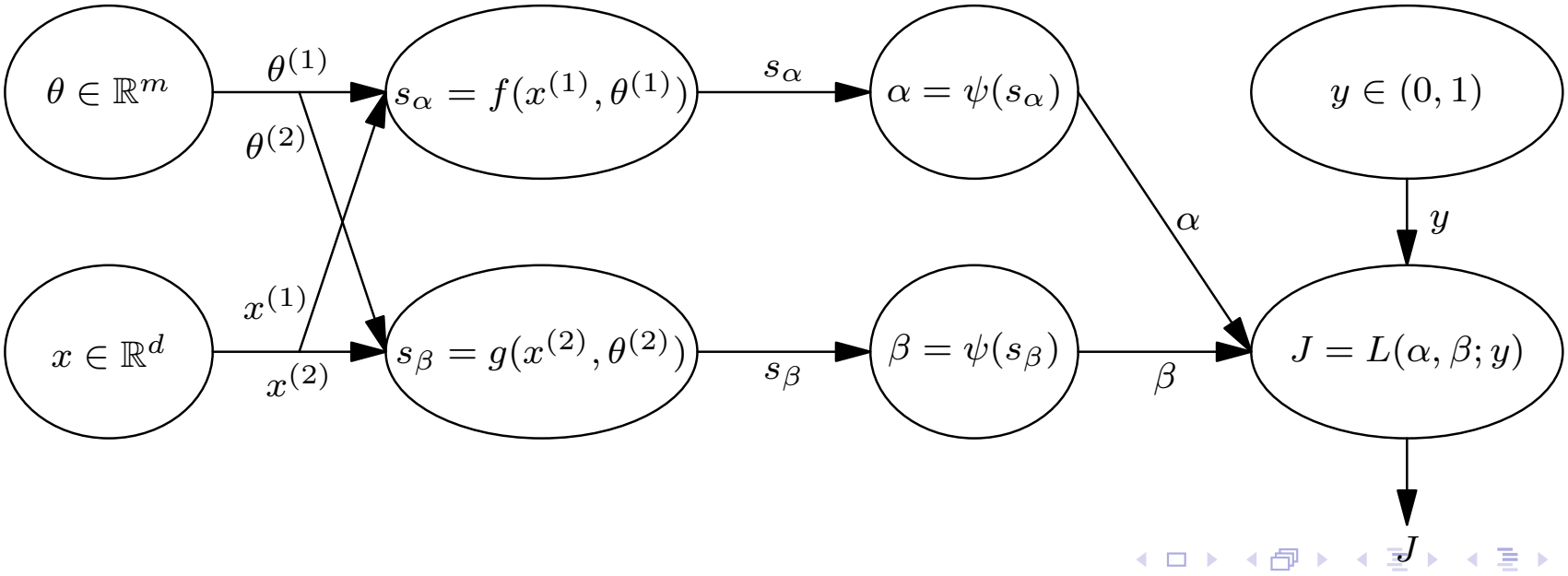
Neural Networks

We have a neural network with one input neuron in the first layer, a hidden layer with n neurons and a output layer with 1 neuron. The hidden layer has $ReLU$ activation functions. You are allowed to use bias. Explain if the following functions can be represented in this architecture and if yes, give a value for n and set of weights for which this is true.

1. $f(x) = |x|$

Backprop

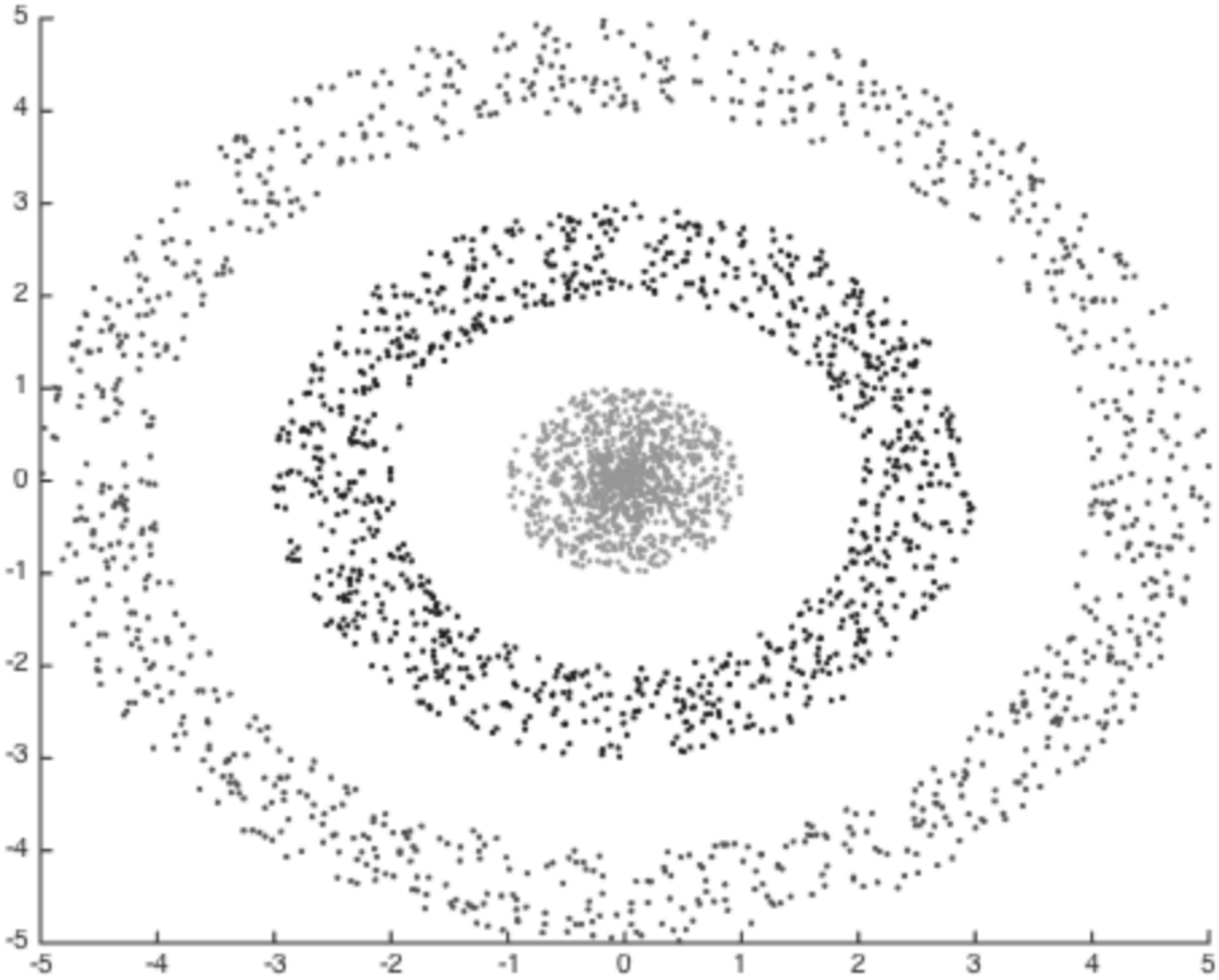
Suppose $f(x;\theta)$ and $g(x;\theta)$ are differentiable with respect to θ. For any example (x, y), the log-likelihood function is $J = L(\theta; x, y)$. Give an expression for the scalar value $\frac{\partial J}{\partial \theta_i}$ in terms of the "local" partial derivatives at each node. That is, you may assume you know the partial derivative of the output of any node with respect to each of its scalar inputs. For example, you may write your expression in terms of $\frac{\partial J}{\partial \alpha}, \frac{\partial J}{\partial \beta}, \frac{\partial \alpha}{\partial s_\alpha}, \frac{\partial s_\alpha}{\partial \theta_i^{(1)}}$, etc. Note that, as discussed in lecture, $\theta^{(1)}$ and $\theta^{(2)}$ represent identical copies of θ, and similarly for $x^{(1)}$ and $x^{(2)}$.



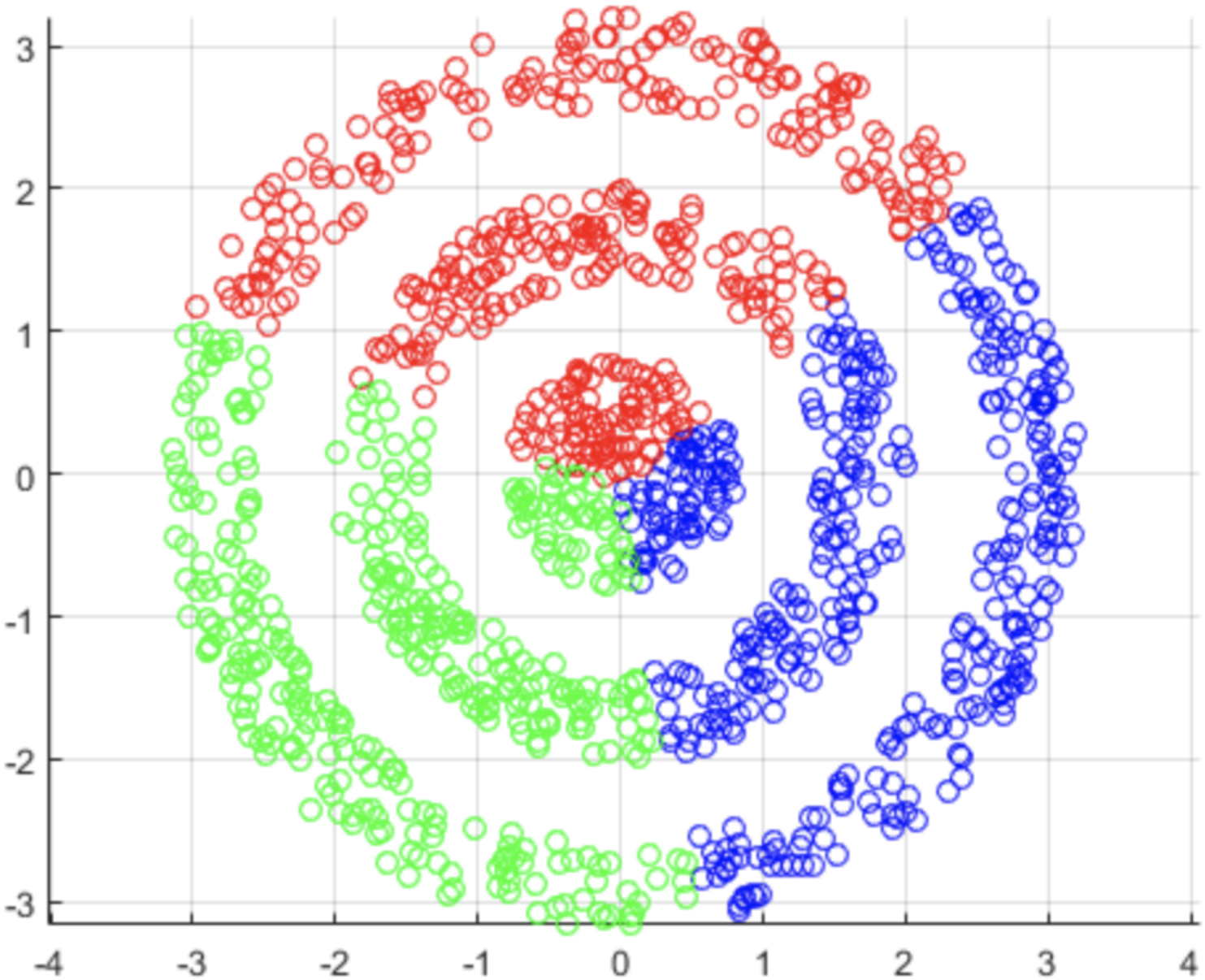
Backprop solution

$$\frac{\partial J}{\partial \theta_i} = \frac{\partial J}{\partial \alpha} \frac{\partial \alpha}{\partial s_\alpha} \frac{\partial s_\alpha}{\partial \theta_i^{(1)}} + \frac{\partial J}{\partial \beta} \frac{\partial \beta}{\partial s_\beta} \frac{\partial s_\beta}{\partial \theta_i^{(2)}}$$

KMeans



KMeans Solution



Mixture Models

Suppose we have a latent variable $z \in \{1, 2, 3\}$ and an observed variable $x \in (0, \infty)$ generated as follows:

$$z \sim \text{Categorical}(\pi_1, \pi_2, \pi_3)$$

$$x \mid z \sim \text{Gamma}(2, \beta_z),$$

where $(\beta_1, \beta_2, \beta_3) \in (0, \infty)^3$, and $\text{Gamma}(2, \beta)$ is supported on $(0, \infty)$ and has density $p(x) = \beta^2 x e^{-\beta x}$. Suppose we know that $\beta_1 = 1, \beta_2 = 2, \beta_3 = 4$. Give an explicit expression for $p(z = 1|x = 1)$ in terms of the unknown parameters π_1, π_2, π_3.

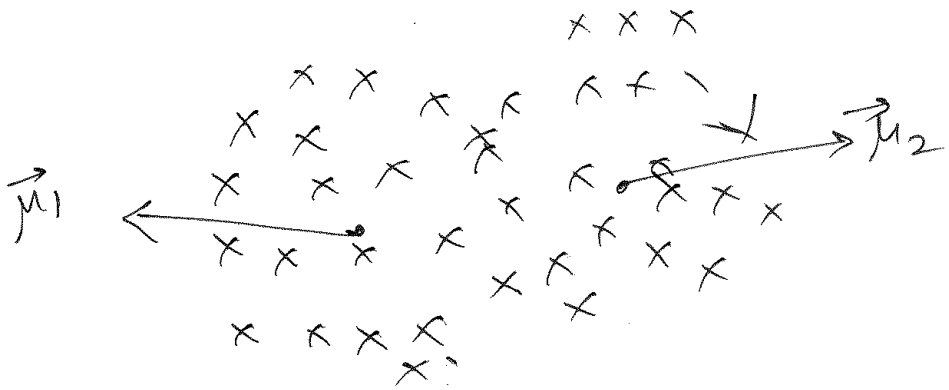
Solutions

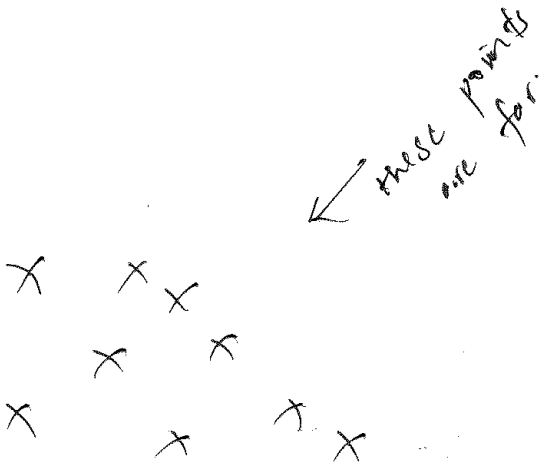
$$p(z=1|x=1) \propto p(z=1|x=1)p(z=1) = \pi_1 e^{-1}$$
$$p(z=2|x=1) \propto p(z=2|x=1)p(z=2) = \pi_2 4e^{-2}$$
$$p(z=3|x=1) \propto p(z=3|x=1)p(z=3) = \pi_3 16e^{-4}$$

$$p(z=1|x=1) = \frac{\pi_1 e^{-1}}{\pi_1 e^{-1} + \pi_2 4e^{-2} + \pi_3 16e^{-4}}$$

GMM for Density Estimation

Assume that clusters 1 and 2 have diagonal co variance matrix $\sigma^2 I$. When is $f_D(\mu_1) > f_D(\frac{\mu_1+\mu_2}{2}$?





we fit a GMM using K latent variables

K classes in total.
or
latent variable.

$\pi_1 = \pi_2 = \ldots = \pi_K$.

let f_D represent density of data

$f_D(\mu_1) > f_D(\frac{\mu_1+\mu_2}{2})$