Recitation 14

Final Review - Questions

Vishakh

CDS

May 1, 2022

Bayesian

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) 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.

Bayesian - Continued

1. Give an expression for the likelihood function $L_D(\theta)$ 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?

Boostrap

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

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.

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 $f_i \in \mathcal{H}_G$

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 $\epsilon > 0$, there exists $m \in \{1, 2, 3, \ldots\}$, $a = (a_1, \ldots, a_m) \in \mathbb{R}^m$, $b = (b_1, \ldots, b_m) \in \mathbb{R}^m$, 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 $x \in [0,1]^d$.

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.