Machine Learning in High Energy Physics Summer School

Practical Session: Bayesian methods

July 4th 2019

First part: Bayesian reasoning

1. During medical checkup, one of the tests indicates a serious disease. The test has high accuracy 99% (probability of true positive is 99%, probability of true negative is 99%). However, the disease is quite rare, and only one person in 10000 is affected. Calculate the probability that the examined person has the disease.

2. Let $X = \{x_1, \ldots, x_N\}$ be N independent dice rolls. For brevity, we denote the number of times a dice comes up as face $k \in \{1, \ldots, K\}$ as $N_k = \sum_{n=1}^{N} \mathbb{I}(x_n = k)$. With this notation the likelihood has the form

$$p(X \mid \theta) = \prod_{k=1}^{K} \theta_k^{N_k}, \tag{1}$$

   where θ_k is the probability of outcome k. Compute the maximum likelihood estimate for $\theta = (\theta_1, \ldots, \theta_K)$. Do not forget that $\theta \in S_K$, i.e. $\sum_{k=1}^{K} \theta_k = 1$ and $\theta_k \geq 0$ for $k = 1, \ldots, K$.

3. The conjugate prior distribution for multinomial likelihood defined in Eq. 1 is the Dirichlet distribution:

$$\mathrm{Dir}(\theta|\alpha) = \frac{1}{B(\alpha_1, \ldots, \alpha_K)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}, \quad \theta \in S_K$$

   where $\alpha_k > 0$ and $B(\alpha_1, \ldots, \alpha_K) = \int_{S_K} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \mathrm{d}\theta$ is the normalizing constant, also known as the multivariate Beta function. Check that the Dirichlet distribution is indeed the conjugate distribution for multinomial likelihood. Train the model by computing the posterior $p(\theta|X, \alpha)$. Then, compute the posterior predictive $p(x_{N+1} = k|X, \alpha) = \int_{S_K} p(x_{N+1} = k|\theta) p(\theta|X, \alpha) \mathrm{d}\theta$.

   To simplify the answer, you may use the following expression for the multivariate Beta function

$$B(\alpha_1, \ldots, \alpha_K) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K} \alpha_k)}$$

   and the multiplicative property of the Gamma function $\Gamma(\alpha + 1) = \alpha \Gamma(\alpha)$.

Second part: Variational Inference

4. Consider a clustering problem: we have a dataset $X = \{x_i\}_{i=1}^{n}$, where $x_i \in \mathbb{R}^d$ — feature vectors, and want to group these objects into K clusters. To do so we assume that points from the dataset were generated from a Gaussian Mixture Model with K components. For each object x_i we establish additional latent variable z_i which denotes the index of the gaussian from which i-th object was generated. For convenience we use binary vector notation for z_i:

$$z_i \in \{0, 1\}^K, \quad \sum_{k=1}^{K} z_{ik} = 1$$

For Gaussian Mixture Model we use the following probabilistic model:

$$p(X, Z, \pi|\mu, \Sigma) = p(\pi) \prod_{i=1}^{n} p(z_i|\pi) p(x_i|z_i, \mu, \Sigma)$$

$$p(\pi) = \mathrm{Dir}(\pi|\alpha)$$

$$p(z_i|\pi) = \prod_{k=1}^{K} \pi_k^{z_{ik}}$$

$$p(x_i|z_i, \mu, \Sigma) = \prod_{k=1}^{K} \mathcal{N}(x_i|\mu_k, \Sigma_k)^{z_{ik}}$$

Latent variable $\pi = (\pi_1, \ldots, \pi_K)$ denotes probabilities of Gaussian components in the mixture and is restricted to simplex: $\sum_{k=1}^{K} \pi_k = 1$ and $\pi_k \geq 0$ for $k = 1, \ldots, K$. μ and Σ contain parameters of Gaussian components and can be trained with EM-algorithm.

The task is to compute the posterior distribution $p(Z, \pi|X, \mu, \Sigma)$. Firstly, show that likelihood $p(X|Z, \pi, \mu, \Sigma)$ and prior $p(Z, \pi|\mu, \Sigma)$ are not conjugate. Then show that there is a conditional conjugacy of likelihood and priors $p(Z|\pi, \mu, \Sigma)$ and $p(\pi|Z, \mu, \Sigma)$.

Using the following factorized approximation of the posterior:

$$p(Z, \pi|X, \mu, \Sigma) \approx q(Z, \pi) = q(Z)q(\pi)$$

write down update rules for $q(Z)$ and $q(\pi)$ for Variational inference.