THU-70250043, Pattern Recognition (Spring 2017) **Homework: 1**

Bayesian Methods

Lecturer: Changshui Zhang zcs@mail.tsinghua.edu.cn

Student:

MLE and MAP

Maximum Likelihood Estimation (MLE) and Maximum A Posterior (MAP) are two basic principles for learning parametric distributions. In this problem you will derive the MLE and the MAP estimates for some widely-used distributions.

Before stating the problems, we first give a brief review of MLE and MAP. Suppose we consider a family of distributions (c.d.f or p.m.f.) $F := \{f(x|\theta) : \theta \in \Theta\}$, where x denotes the random vector, θ denotes a vector of parameters, and Θ denotes the set of all possible values of θ. Given a set $\{x_1, x_2, ..., x_n\}$ of sample points independently drawn from some $f^* \in F$, or equivalently some $f(x|\theta^*)$ such that $\theta^* \in \Theta$, we want to obtain an estimate of the value of θ^*. Recall that in the case of an independently and identically distributed(i.i.d.) sample the log-likelihood function is in the following form

$$l(\theta) = \sum_{i=1}^{n} \log f(x_i|\theta), \tag{1}$$

which is a function of θ under some fixed sample $\{x_1, x_2, ..., x_n\}$. The MLE estimate $\hat{\theta}_{mle}$ is then defined as follows:

- $\hat{\theta}_{mle} \in \Theta$,
- $\forall \theta \in \Theta$, $l(\theta) \leq l(\hat{\theta}_{mle})$.

If we have access to some prior distribution $P(\theta)$ over Θ, be it from past experiences or domain knowledge or simply belief, we can think about the posterior distribution over Θ:

$$q(\theta) := \frac{(\prod_{i=1}^{n} f(x_t|\theta))\, p(\theta)}{z(x_1, x_2, ..., x_n)}, \tag{2}$$

where

$$z(x_1, x_2, ..., x_n) := \int_{\Theta} \left(\prod_{i=1}^{n} f(x_t|\theta) \right) p(\theta) d\theta. \tag{3}$$

The MAP estimate $\hat{\theta}_{map}$ is then defined as follows:

- $\hat{\theta}_{map} \in \Theta$,
- $\forall \theta \in \Theta$, $q(\theta) \leq q(\hat{\theta}_{map})$, or equivalently,

$$l(\theta) + \log p(\theta) \leq l(\hat{\theta}_{map}) + \log p(\hat{\theta}_{map}). \tag{4}$$

1. The Poisson distribution is useful for modeling the number of events occurring within a unit time, such as the number of packets arrived at some server per minute. The probability mass function of a Poisson distribution is as follows:

$$P(k|\lambda) := \frac{\lambda^k e^{-\lambda}}{k!}, \tag{5}$$

where $\lambda > 0$ is the parameter of the distribution and $k \in \{0, 1, 2, ...\}$ is the discrete random variable modeling the number of events encountered per unit time.

1.1. Let $\{k_1, k_2, ..., k_n\}$ be an i.i.d. sample drawn from a Poisson distribution with parameter λ. Derive the MLE estimate $\hat{\lambda}_{mle}$ of λ based on this sample.

1.2. Let K be a random variable following a Poisson distribution with parameter λ. Derive its mean E[K] and variance Var[K]. Since $\hat{\lambda}_{mle}$ depends on the sample used for estimation, it is also a random variable. Derive the mean and the variance of $\hat{\lambda}_{mle}$, and compare them with E[K] and Var[K]. What do you find?

1.3. Suppose you believe the Gamma distribution

$$p(\lambda) := \frac{\lambda^{\alpha-1} e^{-\lambda/\beta}}{\Gamma(\alpha)\beta^\alpha}, \tag{6}$$

is a good prior for λ, where $\Gamma(\cdot)$ is the Gamma function, and you also know the values of the two hyper-parameters $\alpha > 1$ and $\beta > 0$. Derive the MAP estimation $\hat{\lambda}_{map}$.

1.4. What happens to $\hat{\lambda}_{map}$ when the sample size n goes to zero or infinity? How do they relate to the prior distribution and $\hat{\lambda}_{mle}$?

2. The density function of a $p-$dimensional Gaussian distribution is as follows,

$$N(x|\mu, \Lambda^{-1}) := \frac{\exp(-\frac{1}{2})(x-\mu)^T \Lambda (x-\mu)}{(2\pi)^{p/2}\sqrt{|\Lambda^{-1}|}}, \tag{7}$$

where Λ is the inverse of the covariance matrix, or the so-called precision matrix. Let $\{x_1, x_2, ..., x_n\}$ be an i.i.d. sample from a $p-$dimensional Gaussian distribution.

2.1. Suppose that $n \gg p$. Derive the MLE estimates $\hat{\mu}_{mle}$ and $\hat{\Lambda}_{mle}$.

2.2. Suppose you believe the Gaussian-Wishart prior defined as

$$gw(\mu, \Lambda) := \ N(\mu|\mu_0, (s\Lambda)^{-1}) W(\Lambda|V, v) \tag{8}$$

is a good prior for μ and Λ, where

$$W(\Lambda|V, v) := \frac{|\Lambda|^{(v-p-1)/2}}{Z(V, v)} \exp\left(-\frac{tr(V^{-1}\Lambda)}{2}\right) \tag{9}$$

with $tr(\cdot)$ being the trace of a square matrix and $Z(V, v)$ the normalization term. You also know the values of the hyper-parameters $\mu_0 \in \mathbb{R}^p, s > 0, v > p + 1$, and $V \in \mathbb{R}^{p\times p}$ being positive definite. Derive the MAP estimates $\hat{\mu}_{map}$ and $\hat{\Lambda}_{map}$.

2.3. Again, what happens to $\hat{\mu}_{map}$ and $\hat{\Lambda}_{map}$ when n goes to zero or infinity? How do they relate to the prior distribution and the MLE estimates?

3. It is known that MLEs do not always exist. Even if they do, they may not be unique.

3.1. Give an example where MLEs do not exist.

3.2. Give an example where MLEs exist but are not unique. Please specify the family of distributions being considered, and the kind of samples from which multiple MLEs can be found.

3.3. By finding the two examples as described above, hopefully you have gained some intuition on the properties of the log-likelihood that are crucial to the existence and uniqueness of MLE. What are those properties?

4. Consider a training data of N i.i.d. (independently and identically distribute) observations, $\boldsymbol{X} = \{x_1, x_2, ..., x_N\}$ with corresponding N target values $\boldsymbol{T} = \{t_1, t_2, ..., t_N\}$.

We want to fit these observations into some model

$$t = y(x, \boldsymbol{w}) + \epsilon \tag{10}$$

where $\boldsymbol{w}$ is the model parameters and ϵ is some error term.

4.1 To find $\boldsymbol{w}$, we can minimize the sum of square error

$$E(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \boldsymbol{w}) - t_n\}^2 \tag{11}$$

Now suppose we believe that the distribution of error term ϵ is gaussian

$$p(\epsilon|\beta) = \mathcal{N}(\epsilon|0, \beta^{-1}) \tag{12}$$

where $\beta = \frac{1}{\sigma^2}$ is the inverse of variance. Using the property of gaussian distribution, we have

$$p(t|x, \boldsymbol{w}, \beta) = \mathcal{N}(t|y(x, \boldsymbol{w}), \beta^{-1}) \tag{13}$$

Under this assumption, the likelihood function is given by

$$p(\boldsymbol{T}|\boldsymbol{X}, \boldsymbol{w}, \beta) = \prod_{n=1}^{N}\mathcal{N}(t_n|y(x_n, \boldsymbol{w}), \beta^{-1}) \tag{14}$$

Show that the problem of finding the maximum likelihood (ML) solution for $\boldsymbol{w}$ is equivalent to the problem of minimizing the sum of square error (11).

4.2 In order to avoid overfitting, we often add a weight decay term to (11)

$$E(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \boldsymbol{w}) - t_n\}^2 + \frac{1}{2}||\boldsymbol{w}||^2 \tag{15}$$

On the other hand, we believe that $\boldsymbol{w}$ has a prior distribution of

$$p(\boldsymbol{w}|\alpha) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \alpha^{-1}\boldsymbol{I}) \tag{16}$$

Using Bayes theorem, the posterior distribution for $\boldsymbol{w}$ is proportional to the product of the prior distribution and the likelihood function

$$p(\boldsymbol{w}|\boldsymbol{X}, \boldsymbol{T}, \alpha, \beta) \propto p(\boldsymbol{T}|\boldsymbol{X}, \boldsymbol{w}, \beta)p(\boldsymbol{w}|\alpha) \tag{17}$$

Show that the problem of finding the maximum of the posterior (MAP) solution for $\boldsymbol{w}$ is equivalent to the problem of minimizing (15).

Naive Bayes

1. Considers the learning function $X \to Y$, where class label $Y \in \{T, F\}$, $X = \langle X_1, X_2, ..., X_n \rangle$ where X_1 is a boolean variable and $\{X_2, ..., X_n\}$ are continuous variables. Assume that for each continuous X_i, $P(X_i|Y = y)$ follows a Gaussian distribution. List and give the total number of the parameters that you would need to estimate in order to classify a future example using a Naive Bayes classifier. Give the formula for computing $P(Y|X)$ in terms of these parameters and feature variables X_i.

2. Consider a simple learning problem of determining whether Alice and Bob from CA will go to hiking or not $Y : Hike \in \{T, F\}$ given the weather conditions $X_1 : Sunny \in \{T, F\}$ and $X_2 : Windy \in \{T, F\}$ by a Naive Bayes classifier. Using training data, we estimated the parameters $P(Hike) = 0.5$, $P(Sunny|Hike) = 0.9$, $P(Windy|\neg Hike) = 0.8$, $P(Windy|Hike) = 0.3$ and $P(Sunny|\neg Hike) = 0.4$. Assume that the true distribution of X_1, X_2, and Y satisfies the Naive Bayes assumption of conditional independence with the above parameters.

2.1. Assume Sunny and Windy are truly independent given Hike. Write down the Naive Bayes decision rule for this problem using both attributes Sunny and Windy.

2.2. Given the decision rule above, what is the expected error rate of the Naive Bayes classifier? (The expected error rate is the probability that each class generates an observation where the decision rule is incorrect.)

2.3. What is the joint probability that Alice and Bob go to hiking and the weather is sunny and windy, that is $P(Sunny, Windy, Hike)$?

2.4. Next, suppose that we gather more information about weather conditions and introduce a new feature denoting whether the weather is X_3: Rainy or not. Assume that each day the weather in CA can be either Rainy or Sunny. That is, it can not be both Sunny and Rainy (similarly, it can not be $\neg Sunny$ and $\neg Rainy$). In the above new case, are any of the Naive Bayes assumptions violated? Why (not)? What is the joint probability that Alice and Bob go to hiking and the weather is sunny, windy and not rainy, that is $P(Sunny, Windy, \neg Rainy, Hike)$?

2.5. What is the expected error rate when the Naive Bayes classifier uses all three attributes? Does the performance of Naive Bayes improve by observing the new attribute Rainy? Explain why.

Programming

In this problem you will implement Naive Bayes and Logistic Regression, then compare their performance on a document classification task. The data for this task is taken from the 20 Newsgroups data set, and is available from the attached zip file. The included README.txt describes the data set and file format.

Our Naive Bayes model will use the bag-of-words assumption. This model assumes that each word in a document is drawn independently from a multinomial distribution over possible words. (A multinomial distribution is a generalization of a Bernoulli distribution to multiple values.) Although this model ignores the ordering of words in a document, it works surprisingly well for a number of tasks. We number the words in our vocabulary from 1 to m, where m is the total number of distinct words in all of the documents. Documents from class y are drawn from a class-specific multinomial distribution parameterized by θ_y. θ_y is a vector, where $\theta_{y,i}$ is the probability of drawing word i and $\sum_{i=1}^{m} \theta_{y,i} = 1$. Therefore, the class-conditional probability of drawing document x from our Naive Bayes model is $P(X = x|Y = y) = \prod_{i=1}^{m} (\theta_{y,i})^{count_i(x)}$, where $count_i(x)$ is the number of times word i appears in x.

1. Provide high-level descriptions of the Naive Bayes and Logistic Regression algorithms. Be sure to describe how to estimate the model parameters and how to classify a new example.

2. Imagine that a certain word is never observed in the training data, but occurs in a test instance. What will happen when our Naive Bayes classifier predicts the probability of the this test instance? Explain why this situation

is undesirable. How to avoid this problem? Will logistic regression have a similar problem? Why or why not?

3. Implement Logistic Regression and Naive Bayes. Use add-one smoothing when estimating the parameters of your Naive Bayes classifier. For logistic regression, we found that a step size around 0.0001 worked well. Train both models on the provided training data and predict the labels of the test data. Report the training and test error of both models. Submit your code along with your homework.

4. Which model performs better on this task? Why do you think this is the case?