RICE UNIVERSITY, DEPARTMENT OF COMPUTER SCIENCE

# Assignment 3, COMP 540

Chen Zeng(cz39), Zhihui Xie(zx18)

February 16, 2018

## 1 MAP and MLE parameter estimation

**Answer1** According to the properties of Bernoulli Distribution, we can get:

$$P(D|\theta) = \theta^{\sum x^{(i)}}(1-\theta)^{m-\sum x^{(i)}}$$

According to the principle of maximum likelihood, we can get:

$$\theta^*_{MLE} = argmax_\theta logP(D|\theta) = argmax_\theta \left(\sum x^{(i)} log\theta + \left(m - \sum x^{(i)}\right) log(1-\theta)\right)$$

Make the derivation of $logP(D|\theta)$ on $\theta$ and make it equal to 0, we can get:

$$\frac{\partial logP(D|\theta)}{\partial\theta} = \frac{\partial\left(\sum x^{(i)} log\theta + \left(m - \sum x^{(i)}\right) log(1-\theta)\right)}{\partial\theta} = \frac{\sum x^{(i)}}{\theta} - \frac{\left(m - \sum x^{(i)}\right)}{1-\theta} = 0$$

We can solve the result:

$$\theta = \frac{\sum x^{(i)}}{m}$$

Therefore, here is the final result:

$$\theta^*_{MLE} = \frac{\sum x^{(i)}}{m}$$

This is the estimate of $\theta$.

**Answer2** Because of:

$$Beta(\theta|a,b) \propto \theta^{a-1}(1-\theta)^{b-1}$$

Therefore:

$$P(\theta|D) \propto P(D|\theta)P(\theta) \propto \left(\theta^{\sum x^{(i)}}(1-\theta)^{m-\sum x^{(i)}}\right)\left(\theta^{a-1}(1-\theta)^{b-1}\right) = \theta^{\sum x^{(i)}+a-1}(1-\theta)^{m+b-\sum x^{(i)}-1}$$

According to the MAP estimation of $\theta$, we can get:

$$\theta_{MAP} = argmax_\theta logP(\theta|D) = argmax_\theta \left(\left(\sum x^{(i)} + a - 1\right) log\theta + \left(m + b - \sum x^{(i)} - 1\right) log(1-\theta)\right)$$

Make the derivation of $logP(\theta|D)$ on $\theta$ and make it equal to 0, we can get:

$$\begin{aligned}\frac{\partial logP(\theta|D)}{\partial\theta} &= \frac{\partial\left(\left(\sum x^{(i)} + a - 1\right) log\theta + \left(m + b - \sum x^{(i)} - 1\right) log(1-\theta)\right)}{\partial\theta} \\ &= \frac{\sum x^{(i)} + a - 1}{\theta} - \frac{m + b - \sum x^{(i)} - 1}{1-\theta} \\ &= 0\end{aligned}$$

We can solve the result:

$$\theta = \frac{a + \sum x^{(i)} - 1}{a + b + m - 2}$$

Therefore, here is the final result:

$$\theta^*_{MAP} = \frac{a + \sum x^{(i)} - 1}{a + b + m - 2}$$

When it's under a uniform prior, $a = b = 1$, we can get:

$$\theta^*_{MAP} = \frac{\sum x^{(i)}}{m}$$

Combing the answer from problem1, we can get:

$$\theta^*_{MLE} = \theta^*_{MAP}$$

Therefore, the MAP and MLE estimates of $\theta$ are the same under a uniform prior.

## 2 Logistic regression and Gaussian Naive Bayes

**Answer1** We can easily get the results from the logistic regression shown as below:

$$P\left(y = 1|x\right) = g\left(\theta^T x\right)$$

$$P\left(y = 0|x\right) = 1 - g\left(\theta^T x\right)$$

They are the expression in terms of the parameter $\theta$ and the sigmoid function.

**Answer2** At first, we define the const $A$ as below:

$$A = \frac{1}{P\left(x, y' = 1\right) + P\left(x, y' = 0\right)} = \frac{1}{P\left(x|y' = 1\right)\gamma + P\left(x|y' = 0\right)\left(1-\gamma\right)}$$

Then, according to the definition and properties of Gaussian distribution and Bernoulli distribution, we have:

$$P\left(y = 1\right) = \gamma$$

$$P\left(y = 0\right) = 1 - \gamma$$

$$P\left(x|y = 1\right) = \prod_{j=1}^{d} \frac{1}{\sqrt{2\pi\sigma_j^2}} exp\left(-\frac{\left(x^{(j)} - \mu_j^1\right)^2}{2\sigma_j^2}\right)$$

$$P\left(x|y = 0\right) = \prod_{j=1}^{d} \frac{1}{\sqrt{2\pi\sigma_j^2}} exp\left(-\frac{\left(x^{(j)} - \mu_j^0\right)^2}{2\sigma_j^2}\right)$$

Therefore, according to the Bayes rule and Gaussian Naive Bayes model, we can get the following results:

$$P\left(y = 1|x\right) = \frac{P\left(x|y = 1\right)P\left(y = 1\right)}{\sum_{y' \in \{0,1\}} P\left(x|y'\right)P\left(y'\right)} = A\gamma \prod_{j=1}^{d} \frac{1}{\sqrt{2\pi\sigma_j^2}} exp\left(-\frac{\left(x^{(j)} - \mu_j^1\right)^2}{2\sigma_j^2}\right)$$

$$P\left(y = 0|x\right) = \frac{P\left(x|y = 0\right)P\left(y = 0\right)}{\sum_{y' \in \{0,1\}} P\left(x|y'\right)P\left(y'\right)} = A\left(1-\gamma\right) \prod_{j=1}^{d} \frac{1}{\sqrt{2\pi\sigma_j^2}} exp\left(-\frac{\left(x^{(j)} - \mu_j^0\right)^2}{2\sigma_j^2}\right)$$

Here const $A$ is defined at the beginning.

**Answer3** Because class 1 and class 0 are equally likely, $\gamma = \frac{1}{2}$. Therefore, to simplify the expression for $P\left(y = 1|x\right)$, we can get:

$$P\left(y = 1|x\right) = \frac{A}{2} \prod_{j=1}^{d} \frac{1}{\sqrt{2\pi\sigma_j^2}} exp\left(-\frac{\left(x^{(j)} - \mu_j^1\right)^2}{2\sigma_j^2}\right)$$

For the definition of const $A$:

$$A = \frac{1}{P\left(x, y' = 1\right) + P\left(x, y' = 0\right)} = \frac{2}{P\left(x|y' = 1\right) + P\left(x|y' = 0\right)}$$

Let's divide $P\left(y=1|x\right)$ with $P\left(y=0|x\right)$, using the formation in problem2, we can get:

$$\begin{aligned}\frac{P\left(y=1|x\right)}{P\left(y=0|x\right)} &= \frac{A\gamma\prod_{j=1}^{d}\frac{1}{\sqrt{2\pi\sigma_j^2}}exp\left(-\frac{\left(x^{(j)}-\mu_j^1\right)^2}{2\sigma_j^2}\right)}{A\left(1-\gamma\right)\prod_{j=1}^{d}\frac{1}{\sqrt{2\pi\sigma_j^2}}exp\left(-\frac{\left(x^{(j)}-\mu_j^0\right)^2}{2\sigma_j^2}\right)} \\ &= exp\left(\sum_{j=1}^{d}\left(-\frac{\left(x^{(j)}-\mu_j^1\right)^2}{2\sigma_j^2}+\frac{\left(x^{(j)}-\mu_j^0\right)^2}{2\sigma_j^2}\right)\right) \\ &= exp\left(\sum_{j=1}^{d}\frac{\left(\mu_j^0\right)^2-\left(\mu_j^1\right)^2}{2\sigma_j^2}+\sum_{j=1}^{d}\frac{x^{(j)}\left(\mu_j^1-\mu_j^0\right)}{\sigma_j^2}\right)\end{aligned}$$

According to the answer of problem1, we have:

$$\frac{P\left(y=1|x\right)}{P\left(y=0|x\right)}=\frac{g\left(\theta^T x\right)}{1-g\left(\theta^T x\right)}=exp\left(\theta^T x\right)$$

These two are equal and comparing them, we can get:

$$exp\left(\sum_{j=1}^{d}\frac{\left(\mu_j^0\right)^2-\left(\mu_j^1\right)^2}{2\sigma_j^2}+\sum_{j=1}^{d}\frac{x^{(j)}\left(\mu_j^1-\mu_j^0\right)}{\sigma_j^2}\right)=exp\left(\theta^T x\right)$$

It means we have the following equation:

$$\sum_{j=1}^{d}\frac{\left(\mu_j^0\right)^2-\left(\mu_j^1\right)^2}{2\sigma_j^2}+\sum_{j=1}^{d}\frac{x^{(j)}\left(\mu_j^1-\mu_j^0\right)}{\sigma_j^2}=\theta^T x$$

Comparing the parameters to $x$, we can get the appropriate parameterization:

$$\theta_0=\sum_{j=1}^{d}\frac{\left(\mu_j^0\right)^2-\left(\mu_j^1\right)^2}{2\sigma_j^2}$$

$$\theta_j=\frac{\mu_j^1-\mu_j^0}{\sigma_j^2}\left(j=1,...,d\right)$$

Therefore, it shows that with appropriate parameterization, $P\left(y=1|x\right)$ for Gaussian Naive Bayes with uniform priors is equivalent to $P\left(y=1|x\right)$ for logistic regression.

## 3 Reject option in classifiers

**Answer1** To calculate the minimum risk, we discuss it in the following two cases.
When $i = 1, 2, ..., C$, we have:

$$\begin{aligned} R(\alpha_i|x) &= \sum_{j=1}^{C} L\left(\alpha_i|y=j\right) P\left(y_j|x\right) \\ &= \lambda_s \sum_{j=1, j\neq i}^{C} P\left(y=j|x\right) \\ &= \lambda_s \left(1 - P\left(y=i|x\right)\right) \end{aligned}$$

When $i = C+1$, we have:

$$R(\alpha_{C+1}|x) = \lambda_r$$

Therefore, if we want to decide $y = j$ to obtain the minimum risk, the following two requirements should be fulfilled:

$$R\left(\alpha_j|x\right) \leq R(\alpha_k|x), \left(j, k = 1, 2, ..., C \& j \neq k\right)$$

$$R\left(\alpha_j|x\right) \leq R(\alpha_{C+1}|x), \left(j = 1, 2, ..., C\right)$$

Combining with the results we get before, we need to solve the following two inequalities:

$$\lambda_s \left(1 - P\left(y=j|x\right)\right) \leq \lambda_s \left(1 - P\left(y=k|x\right)\right)$$

$$\lambda_s \left(1 - P\left(y=j|x\right)\right) \leq \lambda_r$$

Finally we get the results:

$$P\left(y=j|x\right) \geq P\left(y=k|x\right)$$

$$P\left(y=j|x\right) \geq 1 - \frac{\lambda_r}{\lambda_s}$$

Therefore, if it fulfilled the above requirements, we decide $y = j$ to obtain the minimum risk.

**Answer2** When $\lambda_r = 0$, which makes $\lambda_r/\lambda_s = 0$, we do always reject. As it increases but not equal to 1, the relative cost of rejection increases, we are more likely not to reject, but to assign $x$ to class $j$. When $\lambda_r/\lambda_s = 1$, we should assign $x$ to class $j$. Here class $j$ meets with the requirements stated in the last problem.

## 4 Kernelizing k-nearest neighbors

**Answer** According to the definition of the dot product of vector $x_1$ and vector $x_2$ from wikipedia, we have:

$$\langle x_1, x_2 \rangle = x_1 \cdot x_2 = \sum_{i=1}^{d} x_1^{(i)} x_2^{(i)}$$

To re-express this classification rule in terms of dot products, we have:

$$d = \sqrt{\|x_1 - x_2\|^2} = \sqrt{\sum_{i=1}^{d} \left(x_1^{(i)} - x_2^{(i)}\right)^2} = \sqrt{\sum_{i=1}^{d} x_1^{(i)} x_1^{(i)} + \sum_{i=1}^{d} x_2^{(i)} x_2^{(i)} - 2\sum_{i=1}^{d} x_1^{(i)} x_2^{(i)}} = \sqrt{\langle x_1, x_1 \rangle + \langle x_2, x_2 \rangle - 2\langle x_1, x_2 \rangle}$$

To make use of the kernel trick to formulate the k-nearest neighbor algorithm, we can just replace the dot products in the above equation with kernel function $K$ and then we can get:

$$d = \sqrt{K(x_1, x_1) + K(x_2, x_2) - 2K(x_1, x_2)}$$

Instead of using the Euclidean distance to calculate the relation between two vectors, we use this function to calculate it. Here is the concrete algorithm:

**Algorithm 1** Kernel Trick Based $K$ Nearest Neighbour

**Inputs:**
  $S = \{s_i\} = \{(x_i, y_i, d_i)\}$
**Initialize:**
  $d_i = 0,\ i = 1, \ldots, n$
**for** i = 1 to n **do**
  $d_i = \sqrt{K(x_i, x_i) + K(x, x) - 2K(x, x_i)}$
**end for**
$S' = \left\{s_j'\right\} = \{\left(x_j', y_j', d_j'\right)\} = (sort(s_i.d_i))[1:k],\ i = 1, \ldots, n,\ j = 1, \ldots, k$
$cnt1 = 0$
$cnt2 = 0$
**for** j = 1 to k **do**
  **if** $s_j'.y_j' = 1$ **then**
    $cnt1 = cnt1 + 1$
  **else**
    $cnt2 = cnt2 + 1$
  **end if**
**end for**
$label = cnt1 > cnt2 ? 1 : -1$

## 5 Constructing kernels

**Answer1** According to the Mercer's theorem, if $k_1$ is a valid kernel, then the Gram matrix $K_1$ whose elements are $k_1\left(x, x'\right), 1 \leq i, j \leq m$ is positive definite. According to the definition

of positive definite, the scalar $z^T K_1 z$ should always be positive for every non zero column vector $z$. Because we have $k\left(x,x^{'}\right), 1 \leq i,j \leq m = ck_1\left(x,x^{'}\right), 1 \leq i,j \leq m$, we can know the Gram matrix for $k$ is $K = cK_1$. So we have $z^T Kz = cz^T K_1 z$. It should also be positive and $K$ should be positive definite if $c > 0$. Therefore, according to the Mercer's theorem, under the case of $c > 0$, $k$ should also be valid kernels.

**Answer2** Here $k\left(x,x^{'}\right)$ may not be valid kernel and it depends on $f(x)$. For example, if $f(x) = 0$, then $K$ will be a zero matrix and it's not positive definite. Therefore, according to the Mercer's theorem, under this case, $k$ is not a valid kernel.
However, if $f(x)$ is not always equal to 0, $k$ is a valid kernel. Here is the proof:
According to the Mercer's theorem, because $k_1$ is a valid kernel, we have $k_1\left(x,x^{'}\right) = \left\langle \phi(x), \phi\left(x^{'}\right)\right\rangle$.
Therefore, we can make the following transformation:

$$\begin{aligned} k\left(x,x^{'}\right) &= f(x)\, k_1\left(x,x^{'}\right) f\left(x^{'}\right) \\ &= f(x)\, f\left(x^{'}\right)\left\langle \phi(x), \phi\left(x^{'}\right)\right\rangle \\ &= \left\langle f(x)\phi(x), f\left(x^{'}\right)\phi\left(x^{'}\right)\right\rangle \\ &= \left\langle \phi^{'}(x), \phi^{'}\left(x^{'}\right)\right\rangle \end{aligned}$$

Here we define:

$$\phi^{'}(x) = f(x)\phi(x)$$

Therefore, we have:

$$k\left(x,x^{'}\right) = \left\langle \phi^{'}(x), \phi^{'}\left(x^{'}\right)\right\rangle$$

According to the Mercer's theorem, $k$ is a valid kernel

**Answer3** This is similar to answer1. Here $K = K_1 + K_2$ and $z^T Kz = z^T K_1 z + z^T K_2 z > 0$ for every non zero column vector $z$. According to the Mercer's theorem, $K$ is positive definite and therefore $k$ should also be valid kernels.

# 6 One-vs-all logistic regression

**Answer**

# 7 Softmax regression

**Answer 7.1** CIFAR-10 database has 10 labels and $\theta$ matrix is randomly initialized, so $\frac{exp(\theta^{(k)^T} x^{(i)})}{\sum_{j=1}^{K} exp(\theta^{(j)^T} x^{(i)})}$ is approximately 0.1, which means -log($\frac{exp(\theta^{(k)^T} x^{(i)})}{\sum_{j=1}^{K} exp(\theta^{(j)^T} x^{(i)})}$) is -log(0.1). Because $\lambda = 0$, J($\theta$) is about -log(0.1).

| Classes | plane | car | bird | cat | deer | dog | frog | horse | ship | truck |
|---|---|---|---|---|---|---|---|---|---|---|
| OVA | 0.465 | 0.463 | 0.193 | 0.161 | 0.234 | 0.272 | 0.457 | 0.406 | 0.546 | 0.42 |
| Softmax Regression | 0.465 | 0.486 | 0.242 | 0.266 | 0.299 | 0.347 | 0.502 | 0.425 | 0.621 | 0.459 |

### Answer 7.9

Softmax regression claasifier performances better than OVA binary logistic regression.I would recommend Softmax regression claasifier for the CIFAR-10 classification problem. For OVA method, we need to change multiple classes into two classes and we will build K logistic classifiers if it has K classes. In this problem, these 10 classes are mutually exclusive, so softmax regression claasifier would be appropriate and it is faster. What's more, one sample may be classified into serveral classes or non-class.