**Introduction to Machine Learning** **Fall Semester, 2013**

# Lecture 2: October 20

*Lecturer: Eran Halperin* *Scribe: ym*

Some of the material was not presented in class (and is marked with a side line) and is given for completeness.

## 2.1 Topics for Bayesian Inference

In the next two lectures we will study about Bayesian inference. We will cover the following topics:

1. Maximum Likelihood
2. Prior and Posterior distribution
3. Naïve Bayes
4. Expectation Maximization

## 2.2 Motivation

In the previous lecture we considered the problem of clustering using $k$-means algorithm. In that setup we have as input a set of $n$ points and we would like to partition them to $k$ clusters, where cluster $i$ has a center $\mu_i$. We had two tasks:

1. For each cluster $i$ find a center $\mu_i$
2. For each point $x_j$ find to which cluster to assign it.

We set an objective function that we wanted to minimize

$$\min \sum_{i=1}^{k} \sum_{x_j \in S_i} \|\mu_i - x_j\|_2^2$$

A natural question is why use this objective. We can change the norm to another norm, and it will still make sense, but this is a minor change.

One answer might be is that the objective is both natural and computationally attractive (although finding the global minimum is a hard computational problem).

This raises the question:

*What should be the criteria in selecting an objective function?*

We would like it to be computationally attractive, but more important is that we would like it to minimize our error. In order to define what we mean by minimizing the error we need to define the problem more precisely, and this is exactly what we will do today.

At a high level, we have the axiom which is at the core of machine learning:

*Historical data and future data come from the same distribution.*

This implies that it would be natural to define what is this underlying distribution.

## 2.3 The Likelihood function

Assume we observe $n$ coin tosses of a coin with bias $p$ and observe the number of 'heads'. Namely, we observe a Bernoulli random variable with parameters $p$ and $n$, let $X \sim B(n,p)$. Recall that,

$$\Pr[X = m | X \sim B(n,p)] = \binom{n}{m} p^m (1-p)^{n-m}$$

Now assume that the parameter $p$ is unknown. *How can we learn p from the data?*

We will define a likelihood function,

$$L(p; X = m) = \Pr[X = m | X \sim B(n,p)]$$

The likelihood function measures the probability of observing the data ($X = m$) given a model (a value for $p$).

Here is the general methodology of the likelihood function. Assume we have a set of hypotheses to choose from. Normally a hypothesis will be defined by a set of parameters $\theta$. We do not know $\theta$, but we make some observations and get data $D$. The likelihood of $\theta$ is

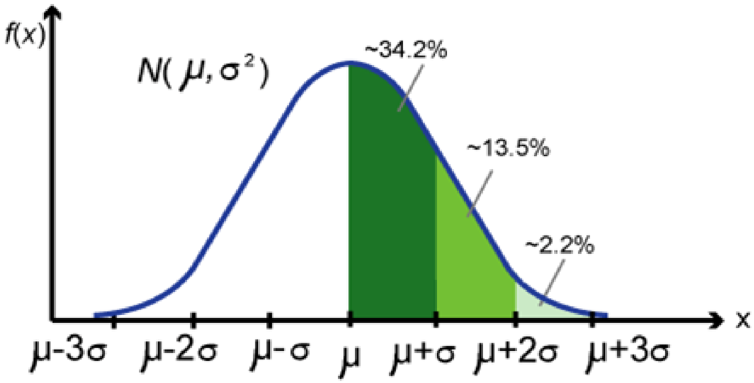
$$L(\theta; D) = Prob(D|\theta).$$

We are interested in the hypothesis that maximizes the likelihood.

Back to our example *What is the p that would maximize the likelihood?*

In order to find the maximizing $p$, we need to compute

$$\begin{aligned} 0 =& \frac{d}{dp} L(p; X = m) \\ =& \binom{n}{m} \left( m p^{m-1} (1-p)^{n-m} - (n-m) p^m (1-p)^{n-m-1} \right) \end{aligned}$$



This is equivalent to

$$mp^{m-1}(1-p)^{n-m} =(n-m)p^{m}(1-p)^{n-m-1}$$

equivalently

$$m(1-p) =(n-m)p$$

and we derive

$$p_{ML} = \frac{m}{n}$$

We need to check that this is indeed the maximum by taking the second derivative. Also, the above assumes that $m \neq 0, n$, which can be derived similarly.

### 2.3.1 Normal distribution

A random variable $X$ is distributed normally with mean $\mu$ and variance $\sigma^2$ is denoted by $X \sim N(\mu, \sigma^2)$. Intuitively, most of the probability are within $\mu \pm 2\sigma$ (95%) and almost all are within $\mu \pm 3\sigma$ (99.8%). Here are the the basic parameters of a univariate Normal distribution.

$$\boxed{\begin{aligned} f(x) &= \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \\ Pr[a \le Z \le b] &= \int_a^b \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx \\ E[Z] &= \mu \\ Var[Z] &= E[(Z-E[Z])^2] = E[Z^2] - E^2[Z] = \sigma^2 \end{aligned}}$$

Assume we have a sample of $n$ i.i.d. drawn from $N(\mu, \sigma^2)$, where $\mu$ and $\sigma$ are unknown. Let $x_1, \ldots, x_n \sim N(\mu, \sigma^2)$.We first write the likelihood function

$$L((\mu,\sigma); x_1, x_2, \ldots, x_n) = \Pr[x_1, \ldots, x_n | \mu, \sigma] = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{1}{2}(\frac{x_i-\mu}{\sigma})^2}$$

We would like to compute the maximum Likelihood estimator, i.e., $(\mu, \sigma)$ that maximize the likelihood.

In order to find the maximum likelihood, we often take the logarithm, which tends to simplify the mathematical expressions. In addition, taking the logarithm may be critical in practice since multiplying large number of small numbers might result in significant numerical problems. Adding the logarithms is much more numerically stable. Back to our problem, we take the logarithm of the likelihood:

$$\ell(\mu,\sigma); x_1, x_2, \ldots, x_n) = \log L((\mu,\sigma); x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} -\frac{1}{2}(\frac{x_i-\mu}{\sigma})^2 - \frac{n}{2}\log 2\pi - n \log \sigma$$

Find the maximum for $\mu$.

$$\begin{aligned} \frac{\partial}{\partial \mu}\ell &= \sum_{i=1}^{n} \frac{1}{\sigma}(\frac{x_i-\mu}{\sigma}) = 0 \\ \sum_{i=1}^{n} x_i &= n \cdot \mu \\ \hat{\mu} &= \frac{\sum_{i=1}^{n} x_i}{n} \end{aligned}$$

The second derivative is $-n/\sigma^2 < 0$, thus we found a maximum.
Note that this value of $\mu$ is independent of the value of $\sigma$, and it is simply the average of

the observations. Now we compute the maximum for $\sigma$, given that $\mu$ is $\hat{\mu}$

$$\frac{\partial}{\partial\sigma}\ell = \sum_{i=1}^{n}\frac{(x_i-\hat{\mu})^2}{\sigma^3} - \frac{n}{\sigma} = 0$$

$$\sum_{i=1}^{n}(x_i-\hat{\mu})^2 = n\cdot\sigma^2$$

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2}$$

where $\hat{\mu}$ was computed before.

### 2.3.2 Example: Uniform distribution

Assume we observe $n$ samples from the uniform distribution over $[0,\theta]$, i.e, $x_i \sim U(0,\theta)$. What is the maximum likelihood estimator.

Let $D = \{x_1,\ldots,x_n\}$. Sort the $x_i$s and let $x_{(1)}, < \ldots < x_{(n)}$ be the sorted order.

- For any $\theta < x_{(n)}$ we have $L(\theta; D) = 0$
- For any $\theta \geq x_{(n)}$ we have $L(\theta; D) = 1/\theta^n$

This implies that the maximum likelihood estimator is $\widehat{\theta}_{ML} = x_{(n)}$.

As a motivation, consider the case we are given the times of conversations in a call center and we would like to learn when is the call center open. We assume that when the call center is open the call appear uniformly during that time.

Note that in this case it is clear that the ML estimator is an underestimate of the true parameter $\theta^*$, since we have that $x_i < \theta^*$ with probability 1. Therefore, in this case we have

$$E[\widehat{\theta}_{ML}] < \theta^*$$

We can take an unbiased estimator, for example take the median or average and multiply it by two. The expected value would be exactly $\theta^*$. The problem might be that we will have a larger variance in the estimator. This is a simple example of the well known tradeoff between bias and variance, where the bias is the expected accuracy, and the variance measures how likely are we to deviate from the expectation.

Here is the computation of the squared error of the estimator. For the average times two we have

$$\begin{aligned} E[(2\frac{\sum_{i=1}^n x_i}{n} - \theta)^2] =& \frac{4}{n^2}E[(\sum_{i=1}^n x_i)^2] - \frac{4\theta}{n}E[\sum_{i=1}^n x_i] + \theta^2 \\ =& \frac{4}{n^2}[n(n-1)\theta^2/4 + n\theta^2/3] - \frac{4\theta}{n}[n\theta/2] + \theta^2 \\ =& \frac{4}{n^2}[-n\theta^2/4 + n\theta^2/3] \\ =& \frac{\theta^2}{12n} \end{aligned}$$

For the maximum we have

$$\begin{aligned} E[(x_{(n)} - \theta)^2] =& E[x_{(n)}^2] - 2\theta E[x_{(n)}] + \theta^2 \\ =& \frac{n}{n+2}\theta^2 - 2\theta\frac{n}{n+1}\theta + \theta^2 \\ =& \frac{-2\theta^2}{n+2} + \frac{2\theta^2}{n+1} \\ =& \frac{2\theta^2}{(n+1)(n+2)} \end{aligned}$$

We can see that the variance in the maximum is significantly smaller for large $n$.

### 2.3.3 Multinomial distribution

Assume we have a subset $H \subset \{0,1\}^k$ of strings of $k$-bits of size $t$. Let $H = \{h_1, \ldots, h_t\}$. We have a distribution $\vec{p} = (p_1, \ldots, p_t)$ over $H$. Each time we draw an $h_i$ using $\vec{p}$. We draw a sample of size $n$ and observe that the word $h_i$ appears $c_i$ times. We would like to estimate the unknown multinomial distribution $\vec{p}$.

We first write the likelihood function:

$$L(p_1, \ldots p_t; c_1, \ldots c_t) = \binom{n}{c_1}\binom{n-c_1}{c_2}\cdots\binom{n-c_1-\cdots-c_{t-1}}{c_t}p_1^{c_1}\cdots p_t^{c_t}$$

We would like to maximize the likelihood given that $\vec{p}$ is a distribution, i.e., $\sum_{i=1}^t p_i = 1$ and $p_i > 0$.

First, consider the log likelihood function,

$$\ell(p_1, \ldots p_t; c_1, \ldots c_t) = \log\left(\binom{n}{c_1}\binom{n-c_1}{c_2}\cdots\binom{n-c_1-\cdots-c_{t-1}}{c_t}\right) + \sum_{i=1}^t c_i \log p_i$$

Since the first term is a constant, we need to solve the following program

$$\max \sum_{i=1}^{t} c_i \log p_i$$
$$\text{s.t.} \ \sum_{i=1}^{t} p_i = 1 \ \ p_i > 0.$$

In order to solve such a constraint optimization problem we can use Lagrange multipliers.

> Consider the following general constraint optimization problem.
>
> $$\max F(x)$$
> $$\text{s.t.} \forall k \quad g_k(x) = 0$$
>
> We can write the following Lagrangian function
>
> $$\mathcal{L} = F(x) + \sum_k \lambda_k g_k(x)$$
>
> The new variables $\lambda_k$ are called the Lagrangian multipliers. We claim that the maximum of the original optimization, $x^*$, has to be a stationary point of the gradient of the Lagrangian. Namely,
>
> $$\nabla \mathcal{L} = 0.$$
>
> However the other direction does not hold, we might have a stationary point which is not the solution to the maximization.
>
> Note that $\frac{\partial}{\partial \lambda_k} \mathcal{L} = g_k(x)$, and therefore we implicitly enforce $g_k(x) = 0$ for all $k$.
>
> For a nice proof and intuition why $\nabla_x F(x) + \sum_k \lambda_k \nabla g_k(x)$, you can look at the Wikipedia (http://en.wikipedia.org/wiki/Lagrange_multiplier).

In the multinomial example we have

$$\mathcal{L}(\vec{p}, \lambda) = \sum_{i=1}^{t} c_i \log p_i + \lambda(1 - \sum_{i=1}^{t} p_i)$$

For the gradient we have

$$\frac{\partial}{\partial p_i} \mathcal{L} = \frac{c_i}{p_i} - \lambda \qquad \frac{\partial}{\partial \lambda} \mathcal{L} = 1 - \sum_{i=1}^{t} p_i$$

Equating the gradient to zero we have

$$p_i = \frac{c_i}{\lambda} \qquad \lambda = \sum_{i=1}^{n} c_i = n$$

In order to show that this is indeed a maximum we need to consider the Hessian matrix $H$ and show that it is negative-definite. It is sufficient to consider the part that involves the original variables $p$ (without the Lagrangian multipliers). The partial derivatives in this case are

$$\begin{aligned} \frac{\partial^2}{\partial p_i \partial p_j}\mathcal{L} =& 0, \\ \frac{\partial^2}{\partial p_i \partial p_i}\mathcal{L} =& -\frac{c_i}{p_i^2}, \end{aligned}$$

Now, for any vector $v$ of $n$ real coordinates,

$$vHv^t = -\sum_{i=1}^{n} \frac{c_i + 1}{p_i^2} v_i^2 < 0$$

which shows that $H$ is negative definite and therefore the solution is indeed a maximum point.

## 2.4 Bayesian Estimators

There is a conceptual question

*What do we believe that the true parameter is?*

Many times we have some knowledge, or intuition, what are better parameters and what are less likely ones. If we have a small sample, we might get a very unrepresentative result when we take the average.

The Bayesian thinking would suggest that the parameter is selected from some distribution, and then the data is generated given this parameter. So we have the following steps:

1. We select $\theta$ from some distribution $\Pr[\theta]$.
2. We generate the data $x_i$ using the model of $\theta$.

A natural task is now, given $\Pr[\theta]$ to compute

$$\Pr[\theta|D]$$

Note that for this question to make sense we must have a distribution over $\theta$.

The distribution over $\theta$, $\Pr[\theta]$, is called the *PRIOR* distribution. Given Bayes rule we like to compute a *POSTERIOR* distribution:

$$\Pr[\theta|D] = \frac{\Pr[D|\theta]\Pr[\theta]}{\Pr[D]}$$

We can now define a new estimator, the *Maximum A Posteriori (MAP)* estimator.

$$\theta_{MAP} = \arg\max_{\theta} \Pr[\theta|D]$$

contrast this with the ML estimator

$$\theta_{ML} = \arg\max_{\theta} \Pr[D|\theta]$$

We can rewrite the MAP to be similar to the ML:

$$\theta_{MAP} = \arg\max_{\theta} \Pr[D|\theta]\Pr[\theta]$$

which follows from Bayes rule (and since $\Pr[D]$ does not influence the maximization).

### 2.4.1 Example: Normal distribution

Assume we have $n$ samples $x_i$ from $N(\mu, 1)$, where $\mu$ is unknown, but it is known that it comes with a prior $\mu \sim N(0, 1)$. We have already computed the ML estimator:

$$\mu_{ML} = \frac{\sum_{i=1}^{n} x_i}{n}$$

We would now like to derive the MAP estimator. We have that

$$\log(\Pr[x_1, \ldots, x_n|\mu]) = -\frac{n}{2}\log(2\pi) - \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2}$$

and

$$\log(\Pr[\mu]) = -\frac{1}{2}\log(2\pi) - \frac{\mu^2}{2}$$

Ignoring the constants, we would like to maximize

$$F = -\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2} - \frac{\mu^2}{2}$$

We can think of this as having an extra point whose value is 0 (for a total of $n+1$ points) and we are minimizing the square-distance, like in $k$-means. We saw that this maximization problem has a solution when $\mu$ is the average. More formally,

$$\frac{d}{d\mu}F = -\mu + \sum_{i=1}^{n}(x_i - \mu) = 0$$

This is indeed a maximum since

$$\frac{d^2}{d^2\mu}F = -1 + \sum_{i=1}^{n}(-1) = -(n+1) < 0$$

The MAP estimator is

$$\widehat{\mu}_{MAP} = \frac{\sum_{i=1}^{n} x_i}{n+1}$$

in contrast to the ML estimator

$$\widehat{\mu}_{ML} = \frac{\sum_{i=1}^{n} x_i}{n}$$

The intuition is that the MAP estimator gives weight to the prior (in this case it can be viewed as an addition phantom sample whose value is the mean).

We now would like to compute the posterior distribution of a normal distribution. We assume as before that $\mu \sim N(0,1)$.

We compute first the probability of $\mu$ given the sample $x_1, \ldots x_n$.
Using Bayes rule we have

$$\Pr[\mu|x_1,\ldots,x_n] = \frac{\Pr[x_1,\ldots,x_n|\mu]\Pr[\mu]}{\Pr[x_1,\ldots,x_n]}$$

We would would like to determine the probability up to a multiplicative constant, as a function of $\mu$, so we can ignore the denominator. We have that

$$\begin{aligned}\Pr[\mu|x_1,\ldots,x_n] \propto & \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(x_i-\mu)^2}\cdot\frac{1}{\sqrt{2\pi}}e^{-\mu^2/2} \\ \propto & e^{-\frac{1}{2}(\sum_{i=1}^{n}(x_i-\mu)^2+\mu^2)} \\ \propto & e^{-\frac{1}{2}(-\sum_{i=1}^{n}2x_i\mu+(n+1)\mu^2)} \\ \propto & exp\left(-\frac{(\mu-\sum_{i=1}^{n}\frac{x_i}{n+1})^2}{2/(n+1)}\right)\end{aligned}$$

In general, consider two distributions $f$ and $g$. If $f(\mu) = Cg(\mu)$, for some constant $C$, since they both sum up to 1 as distributions, we have to have $C = 1$ and the distributions are identical.

We see that the density is proportional to that of $N(\frac{\sum_{i=1}^n x_i}{n+1}, \frac{1}{n+1})$. This implies that it is exactly that distribution!

In the case when we have a family of distributions with parameter $\theta$ and the posterior is in the same family, it is called a *conjugate prior*. The Normal distribution is an example of a conjugate prior.

Computing the posterior distribution gives us much more information than simply the MAP. If we "believe" that our prior is indeed correct, then the posterior includes all the information we have regarding the unknown model. Conceptually, we can think of the posterior of one stage, as a prior of the next stage, and hence we are always only modifying our belief in the model. Also, once we have the posterior distribution, we do not need to maintain any of the samples, since all their information is already included in the posterior.

### 2.4.2 Beta Distribution

Sometimes we can select a prior in such a way that it will be a conjugate prior.

Consider the case of a Bernoulli random variable $X \sim B(n, p)$. *What would be a natural prior?*

Probably a natural prior is a uniform distribution over $[0, 1]$, but the posterior distribution would be far from uniform. We will define a family of distributions, that includes the uniform distribution, and would be the conjugate prior in such a case, called *Beta distributions.*

A Beta distribution has two parameters $\alpha, \beta > 0$. The density function is $f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{C(\alpha,\beta)}$, where $C(\alpha, \beta)$ is a constant whose goal is to normalize the distribution to sum to one. The expected value of the Beta distribution is $\frac{\alpha}{\alpha+\beta}$

Assume that the prior over $p$ is $Beta(\alpha, \beta)$ and we sample $n$ points from $B(n, p)$.

The posterior of $p$ is

$$\Pr[p|X = m, \alpha, \beta] \propto \binom{n}{m} p^m (1-p)^{n-m} \cdot \frac{p^{\alpha-1}(1-p)^{\beta-1}}{C(\alpha, \beta)}$$

where the first part is the likelihood and the second is the prior. This is proportional to

$$\Pr[p|X = m, \alpha, \beta] \propto p^{m+\alpha-1}(1-p)^{n-m+\beta-1}$$

Therefore this is simply a Beta distribution with parameters $m + \alpha$ and $n - m + \beta$.

The MAP is

$$p_{MAP} = \frac{m + \alpha - 1}{n + \alpha + \beta - 2},$$

which is also the mode of the distribution.

Back to the uniform distribution. For $\alpha = 1$ and $\beta = 1$ we have $Beta(1,1)$ is the uniform distribution. Therefore, if we would like to consider a uniform prior, we can select $Beta(1,1)$. The posterior would be the distribution $Beta(m+1, n-m+1)$ and the MAP would be $m/n$.

## 2.5 Naïve Bayes

Will be covered again in the start of lecture three, and will appear in that lecture scribe notes.