Machine Learning	Spring semester, 1996
Lecture 2: Feb. 26, 1997	
Lecturer: Yishay Mansour	*Scribe: Dotan Emanuel*

2.1 Bayesn classification

2.2 Overview

This chapter is based on the Bayesn Model.The Bayesn model is the first model we see regarding Machine learning. In this model we assume we have total information and we will get the optimal results. In this chapter we will get familiar with three basic inference rules:

1. ML - Maximum Likelihood rule.
2. MAP - Maximum A Posteriori rule.
3. BAYES - Posterior Rule.

We will also repeat and make used of the Logaritmic Loss Function interduced last week.

2.3 Bayesn Rule

The Bayesn classification approach is based on the Bayes rule.

$$Pr[A|B] = \frac{Pr[B|A] \cdot Pr[A]}{Pr[B]}$$

and will use it as follows

$$Pr[h|Data] = \frac{Pr[Data|h] \cdot Pr[h]}{Pr[Data]}$$

When:

$$\begin{aligned} Data &= \text{Known information.} \\ h &= \text{Hypotheses regarding the Data distribution.} \end{aligned}$$

Bayes rule will let us get the desired distribution $Pr\left[h|Data\right]$ using the prior distribution $Pr[h]$ and the known Data.

2.4 Example 1: Cancer test

A hospital want to try a new Lab test for Cancer, The company that provide the "Test kit" states that:

1. The test result is Positive in 98% of the cases when the patient does have Cancer.

2. The Test result is Negative in 97% of the cases when the patient does not have Cancer.

3. The Cancer probability in normal population is 1%.

Does this test accurate ?
If a patient gets a positive answer, what are the chances that he really have Cancer ?
Lets calculate those chances ($Pr[Cancer|+]$).

$$\begin{aligned} Pr[Cancer] &= 0.01 \\ Pr[+|Cancer] &= 0.98 \\ Pr[-|!Cancer] &= 0.97 \end{aligned}$$

Using Bayes rule

$$Pr[Cancer|+] = \frac{Pr[+|Cancer] \cdot Pr[Cancer]}{Pr[+]}$$

$$\begin{aligned} Pr[+] &= Pr[Cancer] \cdot Pr[+|Cancer] + Pr[!Cancer] \cdot Pr[+|!Cancer] \\ &= 0.01 \cdot 0.98 \ + \ 0.99 \cdot 0.03 \\ &= 0.0368 \end{aligned}$$

$$Pr[Cancer|+] \ = \ \frac{0.98 \cdot 0.01}{0.0368} \ = \ \frac{0.98}{3.68} \ = \ 0.266$$

We surprisingly find out that this test is useless. we better tell all patient that they dont have cancer (99% of being right) Then count on the test result. And that is although the numbers 97% and 98% looks pretty good when we started. Note that The error comes from the low probability of Cancer in the general population (1%).

2.5 A short reminder of Normal distribution

$$
\begin{aligned}
p[x] &= \tfrac{1}{sqrt[2]2\pi} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)} \\
Prob[a \le X \le b] &= \textstyle\int_a^b p(x)\,dx. \\
E[x] &= \mu \\
Var[x] &= E\left(x - E[x]\right)^2 \\
&= E\left[x^2\right] - E^2[x] \\
&= \sigma^2
\end{aligned}
$$

2.6 Example 2:

We have a set of numbers which follow some unknown normal distribution. We would like to find the distribution parameters (Expectation and Variance). We will assume that the expectation and Variance are both normally distributed with $\mu = 0$ and $\sigma = 1$. By using Bayes rule we can state.

$$
\begin{aligned}
Pr[N(\mu,\sigma)|\left\langle x_1,\ldots,x_m\right\rangle] &= \tfrac{Pr[\langle x_1,\ldots,x_m\rangle|N(\mu,\sigma)]\cdot Pr[N(\mu,\sigma)]}{Pr[\langle x_1,\ldots,x_m\rangle]} \\
Pr[\left\langle x_1,\ldots,x_m\right\rangle|N(\mu,\sigma)] &= \textstyle\prod_{i=1}^m \tfrac{1}{sqrt[2]2\pi} \cdot e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)} \\
Pr[N(\mu,\sigma)] &= \tfrac{1}{sqrt[2]2\pi\sigma} \cdot e^{-\frac{1}{2}\mu^2}
\end{aligned}
$$

$Pr[\langle x_1,\ldots,x_m\rangle]$ is a normalzing factor.
We now have three different approaches:

2.6.1 *Maximum Likelihood*

We aim to choose the hypothesis which best explains the sample, independent of the prior over the hypothesis space, i.e. the hypothesis which maximizes the likelihood consistent with the sample. $\max_{h_i \in \mathcal{H}} \Pr[D|h]$ where $D =$ Data.
In our case $\max_{\mu,\sigma} Pr[\langle x_1,\ldots,x_m\rangle|N(\mu,\sigma)]$.

$$
Pr[\left\langle x_1,\ldots,x_m\right\rangle|N(\mu,\sigma)] = \prod_{i=1}^m \frac{1}{sqrt[2]2\pi} \cdot e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)}
$$

Take a Log (just to easy the computation).

$$L = \sum_{n=1}^{m} -\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right) - m\log\sigma - \frac{m}{2}\log 2\pi$$

and find the Maximum for μ

$$\frac{\partial L}{\partial \mu} = \sum_{i=1}^{m}\left(\frac{x_i-\mu}{\sigma}\right)\cdot\frac{1}{\sigma} = 0.$$

$$\sum_{i=1}^{m} x_i = m\mu$$

$$\mu = \frac{1}{m}\sum_{i=1}^{m} x_i$$

Note that this value is independent of the value of σ.
And now find the Maximum for σ.

$$\frac{\partial L}{\partial \sigma} = \sum_{i=1}^{m}\left(\frac{(x_i-\mu)^2}{\sigma^3}\right) - \frac{m}{\sigma} = 0$$

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

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

Note, In this calculation we did not use the prior known distribution of μ or σ.

2.6.2 *MAP - Maximum A Posterior*

Now we would like to find the best hypothesis taking into account the known distribution of μ and σ.

We aim to maximize

$$\max_{h_i\in\mathcal{H}} \Pr[h_i|D] = \max_{h_i\in\mathcal{H}} \frac{Pr[D|h_i]\cdot Pr[h_i]}{Pr[D]}$$

And since $Pr[D]$ is constant for all $h_i \in \mathcal{H}$ we can omit it.
how will the result we got in the ML approch change ?.
Note, we will have to get a smaller μ since we now add the fact that most of our hypothesis are around 0 (here our hypothesis are the μ for the Expectation distribution, The Expectation has a normal distribution with $\mu = 0$,$\sigma = 1$, which mean that most of our hypothesis are around 0).

$$Pr[h|d] = \prod_{i=1}^{m}\frac{1}{\sqrt{2\pi}}\cdot e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2}\cdot e^{-\frac{1}{2}\mu^2}\cdot e^{-\frac{1}{2}\sigma^2}$$

Take a log

$$L_{map} = \sum_{n=1}^{m} -\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2 - m\log\sigma - \frac{1}{2}\log 2\pi - \frac{1}{2}\mu^2 - \log\frac{1}{2}\sigma^2 - \log 2\pi$$

$$\frac{\partial L_{map}}{\partial\mu} = \sum_{i=1}^{m}\frac{x_i-\mu}{\sigma^2} - \mu = 0$$

$$\frac{\partial L_{map}}{\partial\sigma} = \sum_{i=1}^{m}\frac{(x_i-\mu)^2}{\sigma^3} - \frac{\mu}{\sigma} - \sigma = 0$$

Now we should maximize both equation simultaneously.

$$(\sigma^2+m)\mu = \sum_{i=1}^{m} x_i \qquad \mu = \frac{1}{\mu+\sigma^2}\sum_{i=1}^{m} x_i$$

$$(\sigma^2+m)\sigma^2 = \sum_{i=1}^{m}(x_i-\mu)^2 \qquad \sigma^2 = \frac{\sum_{i=1}^{m}(x_i-\mu)^2}{\sigma^2+m}$$

We will not solve the equations, but it can be easily seen that μ will be closer to zero then in the ML approch.

2.6.3 *Posterior (Bayes)*

we will calculate new distribution for every y.

$$Pr[y|D] = \sum_{i=1}^{m} Pr[h_i|D]\cdot Pr[h_i=y]$$

And in our case

$$Pr[y] = \int\int Pr[N(\mu,\sigma)|D]\cdot Pr[N(\mu,\sigma)|y]d\mu d\sigma$$

Note that this is simply the average distribution function of all our hypotheses. This is the most accurate yet most complex rule.

2.7 Learning a Concept Family

We are given a Concept Family $\mathcal{H}$. Our information consist of sets $\langle x, f(x)\rangle$,$f\in\mathcal{H}$ unknown function. We will assume that the functions in the family $\mathcal{H}$ are Deterministic function. Pr[h(x)=1] = {1,0} . we will also assume that the process that generate the input is independent of the target function f. That means that the chosen points (x_i) alone contain no information on f (the target function).
For each $h\in\mathcal{H}$ we will calculate $Pr[Data|h]$ when $Data = \langle x_i, b_i\rangle \ldots \langle x_m, b_m\rangle$, $b_i = f(x_i)$.

$$\forall i \;\; b_i \neq h(x_i) \;\;, Pr\left[\langle x_i, b_i\rangle|h\right] = 0 \;\Rightarrow\; Pr[Data|h] = 0$$

And

$$\forall i \;\; b_i = h(x_i), \;\; \forall j \;\; b_j = h(x_j) Pr\left[\langle x_i, b_i\rangle \, |h\right] \;\; = \;\; Pr\left[\langle x_j, b_j\rangle \, |h\right]$$

Those hypothesis ($b_i = h(x_i)$) are called consistent functions and will be marked as $\mathcal{H}\prime$.

Now lets use the 3 method's we have described.

1. Maximum Likelihood
   We just have to choose any of the consistent functions.

2. Maximum Aposterior
   First we have to define the Prior function. We will choose $\mathcal{H}$ to be uniformly distributed. $\forall h \in \mathcal{H} \quad Pr(h) = \frac{1}{|\mathcal{H}\prime|}$.
   In this case
   $$Pr[h|Data] = \frac{Pr[Data|h] \cdot Pr[h]}{Pr[Data]} = \begin{cases} 0 & h \notin \mathcal{H}\prime \\ \frac{1}{|\mathcal{H}\prime|} & h \in \mathcal{H}\prime \end{cases}$$
   So we just have to randomly choose one function out of $\mathcal{H}\prime$.

3. Bayes. here we will build an average of the functions in $\mathcal{H}\prime$.
   $$\forall h \in \mathcal{H} \;\; Pr[h|Data] = \frac{1}{|\mathcal{H}\prime|}$$
   $$B(y) = \sum_{h \in \mathcal{H}} h(y) \cdot \frac{1}{|\mathcal{H}\prime|}$$

2.8 Example 1- Independent Experts

Suppose we are given m functions such that $\forall i \quad \forall x \quad \Pr\left[g_i\left(x\right) = c\left(x\right)\right] = \gamma_i$. $c(x) \in \{0,1\}$
For every instance x we may calculate the values $g_1\left(x\right), \cdots, g_m\left(x\right)$. How do we use this values in order to predict $c(x)$?
We shall calculate the posterior probabilities $\Pr\left[c\left(x\right) = 1 \mid \forall i \quad g_i\left(x\right) = b_i\right]$ and $\Pr\left[c\left(x\right) = 0 \mid \forall i \quad g_i\left(x\right)\right.$
and choose according to the highest probability. Using Bayes Eq.

$$Pr[c = 1 \mid \forall i \quad g_i(x) = b_i] = \frac{Pr[c=1] \cdot Pr[c = 1 \mid \forall i \quad g_i(x) = b_i]}{Pr[\forall i \quad g_i(x) = b_i]}$$

For simplicity, let us assume that the learner doesn't have any a priori knowledge about the value of the target concept, i.e. $\Pr\left[c\left(x\right) = 1\right] = \Pr\left[c\left(x\right) = 0\right] = \frac{1}{2}$. Recall that the g_i functions are independent, then it follows that -

$$\begin{aligned} \Pr\left[\forall i \quad g_i\left(x\right) = b_i \mid c\left(x\right) = 1\right] &= \prod_{i=1}^{m} \Pr\left[g_i\left(x\right) = b_i \mid c\left(x\right) = 1\right] \\ &= \prod_{i=1}^{m} \left(1 - \gamma_i\right) \cdot \left(\frac{\gamma_i}{1 - \gamma_i}\right)^{b_i} \end{aligned}$$

- *For identical* γ_i $(\forall i \quad \gamma_i = \gamma)$ -

$$Pr\left[g_i(x) = 1 \mid c(x) = 1\right] = \gamma_i = \gamma$$
$$Pr\left[g_i(x) = 0 \mid c(x) = 1\right] = 1 - \gamma_i = 1 - \gamma$$

$$\Pr\left[\forall i \quad g_i(x) = b_i \mid c(x) = 1\right] = \gamma^k \cdot (1-\gamma)^{m-k} \tag{2.1}$$
$$\Pr\left[\forall i \quad g_i(x) = b_i \mid c(x) = 0\right] = \gamma^{m-k} \cdot (1-\gamma)^k \tag{2.2}$$

When $k = \sum_{i=1}^{m} b_i$.
We would like to know whice is bigger Eq.2.1 or Eq.2.2.

$$\gamma^k \cdot (1-\gamma)^{m-k} \geq \gamma^{m-k} \cdot (1-\gamma)^k$$
$$\left(\frac{\gamma}{1-\gamma}\right)^k \geq \left(\frac{\gamma}{1-\gamma}\right)^{m-k}$$

$\gamma \geq \frac{1}{2}$ Since those are "experts" they should be at least as good as fliping a coin. $\rightarrow \left(\frac{\gamma}{1-\gamma}\right) \geq 1$.

$$k \geq m - k$$
$$k \geq \frac{m}{2}$$

So we will simply decide be the majorty of experts.

- *In the general case* -

$$\Pr\left[g_i(x) = 1 \mid c_t(x) = 1\right] = \gamma_i$$
$$\Pr\left[g_i(x) = 0 \mid c_t(x) = 1\right] = 1 - \gamma_i$$

These two expressions may be combined to a single one -

$$\Pr\left[g_i(x) = b_i \mid c_t(x) = 1\right] = (1-\gamma_i) \cdot \left(\frac{\gamma_i}{1-\gamma_i}\right)^{b_i}$$

and similarly

$$\Pr\left[g_i(x) = b_i \mid c_t(x) = 0\right] = \gamma_i \cdot \left(\frac{1-\gamma_i}{\gamma_i}\right)^{b_i}$$

Recall that the g_i functions are independent, then it follows that -

$$\begin{aligned}
\Pr\left[\forall i \quad g_i(x) = b_i \mid c_t(x) = 1\right] &= \prod_{i=1}^{m} \Pr\left[g_i(x) = b_i \mid c_t(x) = 1\right] \\
&= \prod_{i=1}^{m} (1-\gamma_i) \cdot \left(\frac{\gamma_i}{1-\gamma_i}\right)^{b_i}
\end{aligned}$$

and the decision rule reduces to

$$\prod_{i=1}^{m}(1-\gamma_i)\cdot\left(\frac{\gamma_i}{1-\gamma_i}\right)^{b_i}\geq\prod_{i=1}^{m}\gamma_i\cdot\left(\frac{1-\gamma_i}{\gamma_i}\right)^{b_i}$$

and hence

$$\begin{aligned}\sum_{i=1}^{m}\log(1-\gamma_i)+\sum_{i=1}^{m}b_i\log\left(\frac{\gamma_i}{1-\gamma_i}\right) &\geq \sum_{i=1}^{m}\log\gamma_i+\sum_{i=1}^{m}b_i\log\left(\frac{1-\gamma_i}{\gamma_i}\right)\\ 2\cdot\sum_{i=1}^{m}b_i\log\left(\frac{\gamma_i}{1-\gamma_i}\right) &\geq \sum_{i=1}^{m}\log\left(\frac{\gamma_i}{1-\gamma_i}\right)\\ \frac{\sum_{i=1}^{m}b_i\log\left(\frac{\gamma_i}{1-\gamma_i}\right)}{\sum_{i=1}^{m}\log\left(\frac{\gamma_i}{1-\gamma_i}\right)} &\geq \frac{1}{2}\end{aligned}$$

assigning $w_i=\log\left(\frac{\gamma_i}{1-\gamma_i}\right)$ (which may interpreted as weight attached to the function g_i) and recall $g_i(x)=b_i$, we may finally obtain the *optimal decision rule*

$$\sum_{i=1}^{m}g_i(x)\frac{w_i}{\sum w_i}\geq\frac{1}{2}$$

The $w_i=\log\left(\frac{\gamma_i}{1-\gamma_i}\right)$ is the weight assign to each Expert. Note that if $\gamma_i=\frac{1}{2}$, then $w_i=0$, and if $\gamma_i\approx 1$ then $w_i\approx\infty$ whice is exactly what we wanted.

2.9 Example 2 - Biased Coin

Assume you wish to estimate the probability p that a coin will come up heads. You flip the coin n times and end up with k heads. We would like to estimate the probability that the next flip will come up head. The probability that out of n flips of coin there will be exactly k heads is

$$\Pr[k\mid p]=\binom{n}{k}p^k(1-p)^{n-k}$$

With the Maximum Likelihood approach, one would choose p as to maximize $Pr[k|p]$. This yields

$$p=\frac{k}{n}$$

Yet this result seems unreasonable when n is small. For example, if you flip the coin only once and get a tail, should you believe that it is *impossible* to get a head on the next toss?

Laplace Rule

Let us suppose a *uniform* prior distribution on p. That is, the prior distribution on all the possible coins is uniform; the density function is $\Pr[p] = 1$, i.e.

$$\Pr[p \leq \theta] = \int_0^\theta dp = \theta$$

We turn to calculate the *posterior*

$$\begin{aligned}
\int_0^1 \Pr[k \mid p] \cdot \Pr[p]\, dp &= \int_0^1 \binom{n}{k} p^k (1-p)^{n-k} dp \\
&= \left[\binom{n}{k} \frac{x^{k+1}}{k+1} \cdot (1-x)^{n-k}\right]_0^1 + \int_0^1 \binom{n}{k} \frac{p^{k+1}}{k+1} (n-k)(1-p)^{n-k-1} dp \\
&= \int_0^1 \binom{n}{k+1} p^{k+1} (1-p)^{n-k-1} dp \\
&= \int_0^1 \Pr[k+1 \mid p] \cdot \Pr[p]\, dp
\end{aligned}$$

where the transition from the second to the third expression is due to the identity

$$\binom{n}{k} \frac{(n-k)}{k+1} = \binom{n}{k+1}$$

Comparing both ends of the above sequence of equalities we realize that all the probabilities are equal, and therefore

$$\int_0^1 \Pr[k \mid p] \cdot \Pr[p]\, dp = \frac{1}{n+1}$$

Intuitivley, it meens that for a random choice of the bias p -
any possible number of heads in a sequance of n coin flips is equally likely. In order to calculate the *posterior expectation* we shall denote by (k, n) the sample of k heads out of n coin flips, and write

- $\Pr[(k,n) \mid p] = p^k (1-p)^{n-k}$
- $\Pr[(k,n)] = \int_0^1 p^k (1-p)^{n-k} dp = \frac{1}{n+1} \cdot \frac{1}{\binom{n}{k}}$

Hence, the *posterior expectation* is

$$E[p \mid (k,n)] = \int_0^1 p \frac{\Pr[(k,n) \mid p] \cdot \Pr[p]}{\Pr[(k,n)]} dp$$

$$= \frac{\int_0^1 p \cdot p^k (1-p)^{n-k} dp}{\frac{1}{n+1} \cdot \frac{1}{\begin{pmatrix} n \\ k \end{pmatrix}}}$$

$$= \frac{\frac{1}{n+2} \cdot \frac{1}{\begin{pmatrix} n+1 \\ k+1 \end{pmatrix}}}{\frac{1}{n+1} \cdot \frac{1}{\begin{pmatrix} n \\ k \end{pmatrix}}}$$

$$= \frac{k+1}{n+2}$$

2.9.1 Loss Function

So far we've examined various approaches for making prediction. We would like to have a criterion in order to decide which approach is preferable. A natural thing to consider is how much we loose by making incorrect prediction; consequently the learner's goal is to minimize that loss. In fact we've already used a $0/1$ loss function which correspond to a correct/incorrect prediction. for example in the On-Line model we were interested in the accumulation of the losses according to this function. A more general loss function may assign any value in the range $[0,1]$ for correct and incorrect predictions and thus, in fact, it determine the cost of using each hypothesis. A main concern would be to avoid choosing inappropriate loss function, i.e. a function which wont assign a minimal cost for using the correct hypothesis.

There are two causes for the loss:

1. *Bayes Risk* - it's the loss we cannot avoid since we bound to have it even if we know the target concept. For example, consider the bias coin problem - even if we knew the bias p we would probably always predict 0 (if $p < \frac{1}{2}$) which, on the average, should result in $p \cdot n$ mistakes.
2. The loss due to incorrect estimation of the target concept.

LogLoss function

A commonly use loss function is the LogLoss which states for the bias coin problem that if the learner guesses that the bias is p then the loss will be

$$\begin{array}{c} \log \frac{1}{p} \text{ when the outcome is 1 (head)} \\ \log \frac{1}{1-p} \text{ when the outcome is 0 (tail)} \end{array}$$

If the true bias is θ then the expected LogLoss is

$$\theta \cdot \log \frac{1}{p} + (1-\theta) \cdot \log \frac{1}{1-p}$$

which attains it's minima when $p = \theta$ (as required)

$$H[\theta] = \theta \cdot \log \frac{1}{\theta} + (1-\theta) \cdot \log \frac{1}{1-\theta}$$

This expression, which is known in the *Information Theory* literature as the *Entropy of* θ, is essentially the Bayes Risk. How far are we from the Bayes Risk when using the guess of p according to the Laplace Rule ?

$$\begin{aligned}
E[\text{LogLoss}] &= \int_0^1 \sum_{n=1}^{T} \sum_{k=1}^{n} \left[\theta \cdot \log \frac{n+2}{k+1} + (1-\theta) \cdot \log \frac{n+2}{n-k+1}\right] \cdot \binom{n}{k} \theta^k (1-\theta)^{n-k} d\theta \\
&= \sum_{n=1}^{T} \sum_{k=1}^{n} \binom{n}{k} \log \frac{n+2}{k+1} \int_0^1 \theta \cdot \theta^k (1-\theta)^{n-k} d\theta + \\
&\quad \sum_{n=1}^{T} \sum_{k=1}^{n} \binom{n}{k} \log \frac{n+2}{n-k+1} \int_0^1 \theta^k (1-\theta)^{n-k} d\theta \\
&= \sum_{n=1}^{T} \sum_{k=1}^{n} \frac{1}{n+1} \frac{k+1}{n+2} \log \frac{n+2}{k+1} + \frac{1}{n+1} \frac{n-k+1}{n+2} \log \frac{n+2}{n-k+1} \\
&= \sum_{n=1}^{T} \sum_{k=1}^{n} \frac{1}{n+1} H\left[\frac{k+1}{n+2}\right]
\end{aligned}$$

$$
\begin{aligned}
&= T \int H[\theta]\, d\theta + \sum_{n=1}^{T} \frac{c}{n} \\
&= \text{Bayes Risk } + O(\log T)
\end{aligned}
$$

Hence, by applying the Laplace Rule, we attained the optimal loss (the Bayes Risk) supplemented by a loss which is logarithmic in the number of coin flips (T).