STA561: Probabilistic machine learning

Introduction: MLE, MAP, Bayesian reasoning (28/8/13)

Lecturer: Barbara Engelhardt *Scribes: K. Ulrich, J. Subramanian, N. Raval, J. O'Hollaren*

1 Classifiers

In this lecture, we introduce and formalize methods for building classifiers, and provide intuitive guidelines for the selection of priors. The classification problem is ubiquitous in the sciences, as commonly we want to infer scientifically meaningful labels based on easily obtainable features of the data.

- X – a set of features (a vector of length p)
- Y – a class label ($y \in \{0, 1\}$).

The random variable X may represent things that we can easily detect. For example, X, could be the result of a test for cancer, an image from a camera, or a distance measurement from an infrared sensor on a mobile robot. The class labels, Y, may not be directly observed. Examples of Y could be a cancer diagnosis, whether or not an object is in an image, or whether the robot believes that it is at its destination.

The relationship between X and Y can be visually described through a graphical model as shown in Figure 1. Node X represents the set of observed features. We wish to classify this sample through these observed features, where the class is represented by the node Y. This is done by training a model: estimating parameters θ using available training data. Note that even after observing data, we will assume that there is a distribution over possible values of θ. There will always be uncertainty about the model parameters. This is covered in later chapters, such as in 5.5 when we discuss hierarchical models.

There are two main types of models for classification that we consider. The models for classification that we will discuss fall into one of these two categories:

- *Generative Classifiers* specify how to generate the data by using a *class conditional density*, $\mathsf{p}(X \mid Y = c, \theta)$, and a *class prior*, $\mathsf{p}(Y = c \mid \theta)$. Generative classifiers are generally very easy to fit data to. They are good at handling unlabeled training data, and they do not require recalibration whenever new classes are added.
- *Discriminative Classifiers* specify how to directly fit the *class posterior*, $\mathsf{p}(Y = c \mid X)$. Their main advantage is that they allow for easier preprocessing and are usually better calibrated in terms of their probability estimates. More details about discriminative classifiers will be given in future lectures.

From here on out in these notes, we are assuming a generative classifier.

Representation of Generative and Discriminative Models

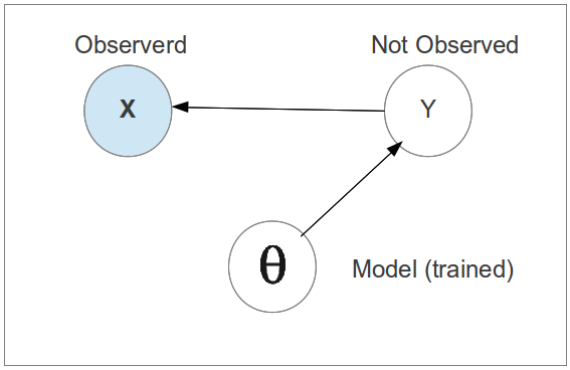

Figure 1: Generative Model: The edge directed from θ to Y (class) illustrates how a priori, Y follows a *class prior* with parameter θ. The edge from Y to X (data) illustrates how these models specify a *class conditional density* for data given class membership. Popular Generative Classifiers include: Naive Byes, Gaussian Mixture Models, Hidden Markov Models etc.

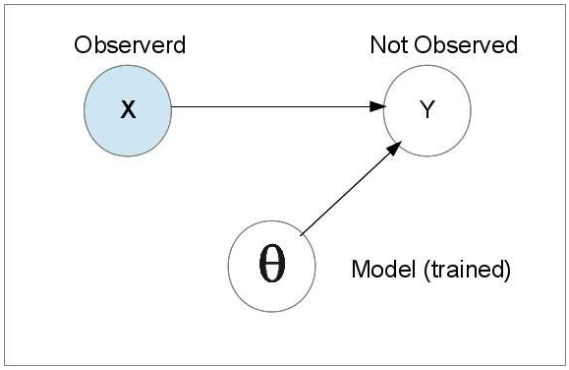

Figure 2: Discriminative Model: Edges are directed from X (data), and θ (parameters) to Y (class) as these methods directly fit the *class posterior* conditional on data. e.g. attempting to infer disease (class) directly from a set of symptoms (data). Popular Discriminative Classifiers include: Logistic Regression, Support Vector Machines, Neural Networks etc.

1.1 Bayes Rule

By combining the definition of conditional probability with the product and sum rules, we can get Bayes Rule. Bayes Rule can be used to obtain a *posterior distribution*:

$$\mathsf{p}(Y = c \mid X, \theta) = \frac{\mathsf{p}(Y = c \mid \theta)\mathsf{p}(X \mid Y = c, \theta)}{\mathsf{p}(X \mid \theta)} \tag{1}$$

where $\mathsf{p}(Y = c \mid \theta)$ is known as the *class prior* and $\mathsf{p}(X \mid Y = c, \theta)$ is known as the *class conditional density*. An important aspect of Bayes rule is the *marginal density* in the denominator. We can redefine the denominator as follows:

$$\mathsf{p}(X \mid \theta) = \sum_{c' \in \mathcal{C}} \mathsf{p}(Y = c' \mid \theta)\mathsf{p}(X \mid Y = c', \theta) \tag{2}$$

Note that this expression is merely a constant for all classes c. Therefore, typically only the numerator of Bayes rule is considered:

$$\mathsf{p}(Y = c \mid X, \theta) \propto \mathsf{p}(Y = c \mid \theta)\mathsf{p}(X \mid Y = c, \theta) \tag{3}$$

This model is *generative* in the sense that we generate a class from $c \sim \mathsf{p}(Y \mid \theta)$ and, given that class, we can generate features that are representative of that class from $c \sim \mathsf{p}(X \mid Y = c, \theta)$. For example, if we generate the class *spam email* then the model will generate a set of features that represent spam such as large numbers of asterisks.

1.2 Posterior Maximization

One of our goals is to find the class assignment which maximizes the posterior. Mathematically, this is:

$$c^* = \arg\max_{c' \in \mathcal{C}} p(Y = c' \mid X, \theta) \tag{4}$$

Before we can compute this, we must specify suitable priors. In the next section, we provide intuition for prior specification by introducing the concept of a *hypothesis space*.

1.3 Why is Bayes Rule Important Here?

In lecture 1 we learned what Bayes Rule is, and in today's lecture we learned how it links together the *prior*, *posterior*, and the *likelihood*. *Bayes Rule allows us to say that the posterior is proportional to the prior times the likelihood.* In plain English this means that Bayes rule allows us to consider the impact of an event having been observed on our belief in various situations. We will now explain these concepts through the use of an example.

2 Hypothesis Space

2.1 Number Game Example

Consider the following example: I'm thinking of a number between 1 and 20 (both inclusive), and I want you to try and guess the number.

In this case, the *hypothesis space* is the set of values that observations can take, $\mathcal{H} = \{1, \ldots, 20\}$. If $N = 4$ values were chosen from this space, the resulting data set might look like the following:

$$\begin{aligned}\mathcal{D}_1 &= \{14, 10, 2, 18\}\\ \mathcal{D}_2 &= \{4, 2, 16, 8\}\\ \mathcal{D}_3 &= \{5, 11, 2, 17\}\\ \mathcal{D}_4 &= \{3, 7, 2, 4\}\end{aligned}$$

Looking at these observations, we may notice something. In particular, $\mathcal{D}_1$ appears to be even numbers, $\mathcal{D}_2$ powers of 2, $\mathcal{D}_3$ prime numbers, and $\mathcal{D}_4$ small numbers. This example illustrates the idea that the data may be generated by a *hidden concept*. Knowing the hidden concept is useful since it reduces the size of the space you have to guess from to get the chosen number.

To simplify calculations, we will assume that the data are sampled uniformly at random (a.k.a *strong sampling assumption*) from a subspace of the hypothesis space. Given this, the *likelihood* of the data for a hypothesis h (assuming $\mathcal{D} \in h$) that restricts the hypothesis space is:

$$\mathsf{p}(\mathcal{D} \mid h \in \mathcal{H}) = \left[\frac{1}{|h|}\right]^N \tag{5}$$

where $|\cdot|$ represents the *size* of the hypothesis space. This brings up the concept of *Occam's razor* – the smallest hypothesis that is consistent with the data is considered to be the best hypothesis. Furthermore, this hypothesis maximizes the likelihood of the observed data.

For example, consider set $\mathcal{D}_2$. We may guess the hidden concept for this set is "powers of 2" or "even numbers". If we consider powers of 2, then there are only 5 numbers within the hypothesis space (i.e. $\mathcal{H} \in \{1, \ldots, 20\}$) that satisfy this rule. On the other hand, if we consider even numbers, there are 10 even numbers in the hypothesis space.

Taking the likelihood ratio of both these concepts under the strong sampling assumption, we get:

$$\begin{aligned}\mathsf{p}(\mathcal{D}|h \in \mathcal{H}) &= \left[\frac{1}{|h|}\right]^N \\ \frac{\mathsf{p}(\mathcal{D}_1|h_1 \in \mathcal{H})}{\mathsf{p}(\mathcal{D}_2|h_2 \in \mathcal{H})} &= \left[\frac{1}{5}\right]^4 / \left[\frac{1}{10}\right]^4 \\ &= 16:1\end{aligned}$$

Thus we can see that the size of hypothesis space has a huge impact on likelihood. The smaller the size of hypothesis space, the higher the likelihood.

Then why not just say that $\mathcal{D}_3$ has a hypothesis $h = \{5, 11, 2, 17\}$? This is, after all, the smallest hypothesis that describes $\mathcal{D}_3$. The reason is that it is conceptually unnatural to choose this hypothesis. Therefore, we want to have a prior which suggests that such unnatural hypotheses are unlikely. This prior can be context-specific and subjective and is denoted as $\mathsf{p}(h \in \mathcal{H})$.

This prior is important because the number of possible hypotheses is huge. In this on-going example, there are 2^{20} possible hypotheses assuming that each number from $1, \ldots, 20$ can either be included or excluded from any hypothesis. Therefore, context specification is extremely important because it helps to eliminate hypotheses that are unlikely. In the above case, it is important to consider who is asking us to choose a number between 1 and 20. Is it a 3 year old child or my professor? If it is a 3 year old, the child may not be familiar with the concept of 'powers of 2' and hence we assign a low value this in the prior. On the other hand, if it is my professor, then we assign higher prior to the same hypothesis. We can choose a prior that has larger probability density on hypotheses that we believe *a priori* to be more likely to be true. This choice of prior (for it is a subjective choice) helps us eliminate hypotheses that are unlikely.

To gain an appreciation for how many hypotheses there are in many real world applications, consider the number of possible gene sequence mutations.

3 Posterior

Using the prior and likelihood described in the previous section (still continuing with the strong sampling assumption), the posterior distribution is written as

$$\mathsf{p}(h \mid \mathcal{D}) \propto \mathsf{p}(\mathcal{D} \mid h)\mathsf{p}(h) \tag{6}$$

$$= \mathsf{p}(h)\left[\frac{\mathbb{1}([\mathcal{D} \in h])}{|h|}\right]^{|\mathcal{D}|} \tag{7}$$

where the *indicator function* $\mathbb{1}(\cdot)$ is defined as 1 if the argument $(\cdot)$ is true and 0 otherwise. The *maximum a posteriori* estimate is the value of h that maximizes this posterior (the posterior mode). This is found according to

$$h^*_{\text{MAP}} = \arg\max_{h' \in \mathcal{H}} \mathsf{p}(h \mid \mathcal{D}) \tag{8}$$

Notice the similarity of the MAP to the MLE (*maximum likelihood estimate*):

$$h^*_{\text{MLE}} = \arg\max_{h' \in \mathcal{H}} \mathsf{p}(\mathcal{D} \mid h) \tag{9}$$

The relationship between the MAP and MLE is illustrated by the following

$$h^*_{\text{MAP}} = \arg\max_{h'} \mathsf{p}(\mathcal{D} \mid h)\mathsf{p}(h) \tag{10}$$

$$= \arg\max_{h'} [\log \mathsf{p}(\mathcal{D} \mid h) + \log \mathsf{p}(h)] \tag{11}$$

$$\propto \arg\max_{h'} [N \log c_1 + c_2] \tag{12}$$

$$\text{as } N \to \infty, \to \arg\max_{h'} \mathsf{p}(\mathcal{D} \mid h) \tag{13}$$

Given infinite observations, the MLE and MAP are equivalent. That is, sufficient data overwhelm the prior. However, when the numbers of observations is small, the prior protects us from incomplete observations.

Both the MLE and MAP are 'consistent', meaning that they converge to the correct hypothesis as the amount of data increases.

3.1 Example

Let us consider a sequence of coin tosses $\mathcal{D} = \{x_1, x_2, \ldots, x_n\}$, where $x_i = 0$ if the i^{th} coin toss was tails, and $x_i = 1$ if it was heads. Here, all x_i are i.i.d. (independent and identically distributed). The probability of heads on a single coin toss is denoted by θ. Hence, $\theta = [0, 1]$. We do not know θ *a priori*. If n_1 is the number of heads and n_0 is the number of tails observed in $n_0 + n_1$ flips, then we can model the likelihood using a Bernoulli as follows:

$$\mathsf{p}(\mathcal{D} \mid \theta) = \prod_{i=1}^{N} \theta^{x_i}(1-\theta)^{(1-x_i)} = \theta^{n_1}(1-\theta)^{n_0} \tag{14}$$

The reasons for using the Bernoulli distribution instead of the binomial are:

1. We are considering only one sequence of heads and tails.
2. Since, data is i.i.d. it is exchangeable. In other words, we can swap any x_i and x_j and it will result in same likelihood. The sequence HTHT is the same as TTHH for our purposes here.
3. $\binom{n}{k}$ is independent of θ.

Sufficient Statistics: Because data sets can be extremely large, it is often important to be able to summarize the data with a finite set of summary statistics. These summary statistics should contain all information that the full data did with respect to estimating the parameters of the given model.

In the coin flip example above, if we have n_1 and n_0 then we no longer need the original sequence of coin flips. This is because they are the *sufficient* for θ. We denote *sufficient statistics* by $\mathcal{S}(\mathcal{D})$. In other words $\mathsf{p}(\theta \mid \mathcal{D}) = \mathsf{p}(\theta \mid \mathcal{S}(\mathcal{D}))$. Because of this property, if we keep $\mathcal{S}(\mathcal{D})$, we can discard all the data.

3.2 Conjugacy

Beta Distribution: Recall that θ is the probability that the coin comes up heads in a single coin flip. We model our parameter θ using beta distribution as follows

$$\mathsf{Beta}(\theta \mid a, b) \propto \theta^{a-1}(1-\theta)^{b-1} \tag{15}$$

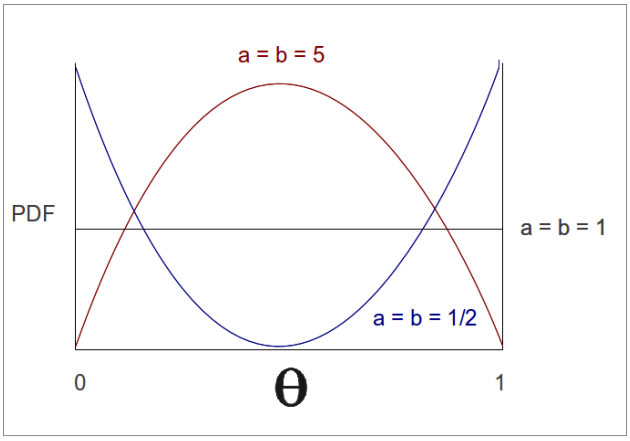

Figure 3: Effects of different parameterizations of the beta distribution.

If we choose $a = b = 1$ such that all θ are equally likely (i.e., an uninformative prior) we get $\mathsf{Beta}(\theta \mid a, b) \propto 1$. On the other hand if we know that our coin is biased towards either heads or tails but do not know which way, we can set $a = b = \frac{1}{2}$ (Refer Figure 3). This way, θ will have high probability close to 0 and 1 and low probability for values in between. If we know that the coin is heavily biased towards heads, we can set $a \gg b$. Here, a and b are called *hyper-parameters*.

The posterior probability $\mathsf{p}(\theta \mid \mathcal{D})$ is given by,

$$\begin{aligned}
\mathsf{p}(\theta \mid D) &\propto \mathsf{p}(\theta \mid a, b)\mathsf{p}(\mathcal{D} \mid \theta) && (16)\\
&\propto \theta^{a-1}(1-\theta)^{b-1}\theta^{n_1}(1-\theta)^{n_0} && (17)\\
&\propto \theta^{n_1+a-1}(1-\theta)^{n_0+b+1} && (18)\\
&\propto \mathsf{Beta}(\theta \mid n_1 + a, n_0 + b) && (19)
\end{aligned}$$

Here, the posterior distribution has the same form as the prior distribution. When this happens, we say that the prior is *conjugate* to the likelihood. The hyper-parameters are effectively *pseudocounts*, corresponding to the number of heads (a) and tails (b) we expect to see in $a + b$ flips of the coin, before collecting any data. For example, we may indicate strong prior belief that the coin is fair by setting $a = b = 100$, or a weak prior belief that the coin is fair by setting $a = b = 2$. Similarly, for a coin biased towards heads we can set $a = 100$ and $b = 1$.

In a nutshell, the posterior is a compromise between the prior and the likelihood. With a conjugate prior, the posterior generally looks like the sum of the hyperparameters and certain data statistics; we will formalize this more when we discuss the exponential family.

3.3 Posterior Mean, Mode, and Variance

In this case, the MAP estimate is given by

$$\hat{\theta}_{\text{MAP}} = \frac{a + n_1 - 1}{a + b + n_0 + n_1 - 2} \tag{20}$$

Using this equation, if we have a and b, we can directly compute the MAP estimate of θ. Also, if $a = b = 1$ then MAP = MLE, which is the mean of the beta distribution.

$$\hat{\theta}_{\text{MLE}} = \frac{n_1}{n_0 + n_1} = \frac{\text{number of heads}}{\text{number of tosses}} \tag{21}$$

Note that, unlike the mean and mode, the median has no closed form. The variance of the posterior is estimated using equation below:

$$\mathsf{Var}[\theta \mid \mathcal{D}] = \frac{(a + n_1)(b + n_0)}{(a + n_1 + b + n_0)^2(a + n_1 + b + n_0 + 1)} \tag{22}$$

If $N \gg a, b$, then

$$\mathsf{Var}[\theta|\mathcal{D}] = \frac{n_1 n_0}{N^3} = \frac{\hat{\theta}_{\text{MLE}}(1 - \hat{\theta}_{\text{MLE}})}{N} \tag{23}$$

The confidence (or, conversely, our uncertainty) in our parameter estimates given our data can be quantified by the *posterior standard deviation* $\sigma \approx \sqrt{\frac{\hat{\theta}_{\text{MLE}}(1-\hat{\theta}_{\text{MLE}})}{N}}$. It shows that uncertainty in an estimate decreases by the rate of $\frac{1}{\sqrt{N}}$. It also shows that a value of $\hat{\theta}_{\text{MLE}} = 0.5$ maximizes uncertainty and a value of $\hat{\theta}_{\text{MLE}} \in \{0, 1\}$ minimizes uncertainty.

4 Posterior Predictive Distribution

We can predict the class of a new sample (e.g., what will be the result of a new coin flip) based on the prior and the data. [**Note: In the book, this is wrongly referred to as the prior predictive distribution**]. In our example, the probability that a new coin flip results in a head is calculated according to

$$\mathsf{p}(X = 1 \mid \mathcal{D}, a, b) = \int_0^1 \mathsf{p}(X = 1 \mid \theta)\mathsf{p}(\theta \mid \mathcal{D})d\theta \tag{24}$$

$$= \int_0^1 \theta \mathsf{Beta}(\theta \mid a, b, n_0, n_1)d\theta \tag{25}$$

$$= \mathbb{E}[\theta \mid \mathcal{D}] \tag{26}$$

$$= \frac{a + n_1}{a + b + n_0 + n_1} \tag{27}$$

Such prediction is incredibly important, as the goal of many statistical models is to predict classes or response vectors based on the observation of new data X^*. Although this example is simple (i.e., the prediction does not depend on any newly observed covariates), the idea of prediction is a powerful one. With this approach, we may predict if new emails are spam, perform image recognition, or recommend movies to users on Netflix.