Chapter 1.2

Paul Scemama

September 2022

1 Probability Theory

1.1 Probability Densities [1.2.1]

Disclaimer: If anyone has trouble with probability theory and grasping the concepts of conditional probability, random variables, etc, I cannot recommend more John Tsitsiklis's course on Applied Probability found here: https://www.youtube.com/watch?v=j9WZyLZCBzs&list=PLmPcD-wiF4Ea_Doghiw3ya6XaLrmGrLUU, while following his and John Bertsekas's book found here: https://ece307.cankaya.edu.tr/uploads/files/introduction%20to%20probability%20(bertsekas,%202nd,%202008).pdf

1.1.1 Conditional probability

Conditional probability gives us a way to speak about uncertain quantities of random variables when we are provided *partial information*. For example, we may have beliefs about likely and unlikely values a random variable x takes. This is captured in its *pdf* – $p(x)$. What if knowing the value of a related random variable y modifies our beliefs about x. Then we could represent this modification as $p(x|y)$. For a given value y^*, $p(x|y = y^*)$ describes a new *pdf* where we now live in a universe where $y = y^*$.

1.1.2 The relationship between conditional and joint distributions

Suppose we have the joint distribution $p(x, y)$ which tells us which values of x are likely to occur with values of y, etc. The joint distribution gives us the most *complete* picture of x and y. The conditional distributions $p(x|y)$ and $p(y|x)$ are *embedded* in the joint. Simply put, the conditional distribution $p(x|y = y^*)$ is a *slice* of the joint distribution, and then normalized so that its probability mass sums to 1, as seen in Figure 1.

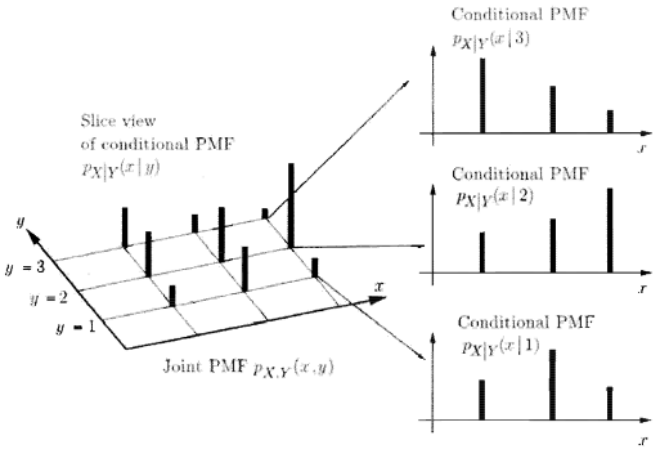

Figure 1: The discrete joint PMF and conditional PMFs derived from the joint.

1.1.3 Sum & Product Rule

The two fundamental rules of probability involve representing a *marginal* distribution in terms of a joint (sum rule) and representing a joint distribution as the product of a conditional and marginal distribution.

sum rule	$p(X) = \sum_Y p(X, Y)$
product rule	$p(X, Y) = p(Y\|X)p(X)$

The *sum rule* tells us that in order to get $p(X)$ from $p(X, Y)$ we consider the probability of X across all possible values of Y and add these probabilities up. The *product rule* tells us how to get from *joint* to conditional or marginal distributions.

1.2 Expectations and Covariance [1.2.2]

1.2.1 Expectations

Consider a function $f(x)$ that depends on a random variable x. x has an underlying distribution that describes how likely it takes on different values. We can take the *expected value* of f under the distribution $p(x)$. This is given by,

Discrete:	$\mathbb{E}[f] = \sum_x f(x)p(x)$
Continuous:	$\mathbb{E}[f] = \int f(x)p(x)dx$

In either case, if we are given a finite number N of points drawn from $p(x)$, then the expectation of f can be approximated as,

$$\mathbb{E}[f] \simeq \frac{1}{N}\sum_{n=1}^{N} f(x_n)$$

We will come across expectations of functions of several variables, in which case Bishop uses the following notation. He uses a subscript to indicate which variable is being averaged over, for instance,

$$\mathbb{E}_x[f(x, y)] = \sum_x f(x, y)p(x)$$

denotes the average of the function $f(x, y)$ with respect to the distribution of x. Note that $\mathbb{E}_x[f(x, y)]$ will be a function of y. What it's saying is: instead of having a function $f(x, y)$ that depends on two random variables, let's instead substitute one of the random variables with its *expected value* (a number) according to its *marginal* distribution. And so this function becomes only dependent on one of the random variables.

We can also consider a *conditional expectation* with respect to a conditional distribution,

$$\mathbb{E}_x[f|y] = \sum_x f(x)p(x|y)$$

First note that this will be a number. A value of y is given; we enter a new universe. Then we look at the distribution of x for that given y. We then take the expected value of $f(x)$ according to this distribution of x. Add visualizations

1.2.2 Variance

The *variance* of $f(x)$ is defined as

$$var[f] = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2]$$

and provides a measure of how much spread or variability there is in $f(x)$ around its mean value $\mathbb{E}[f(x)]$. The variance can also be written as

$$var[f] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2$$

In the special case of $f(x) = x$, we can consider the variance of the variable x itself,

$$var[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2$$

1.2.3 Covariance

The *covariance* expresses the extent to which two random variables *vary* together. It is defined by,

$$\begin{aligned} cov[x, y] &= \mathbb{E}_{x,y}[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}] \\ &= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y]. \end{aligned}$$

If x and y are independent then their covariance vanishes.

1.3 Bayesian probabilities [1.2.3]

1.3.1 Bayesian vs. Frequentist

Frequentist: view probabilities as proportions of long-run frequencies that are the result of repeatable, random events.

- For example, the probability that a coin lands on a head is the proportion of heads that land in infinite repeatable events of flipping the coin.

Bayesian: view probabilities as representations of our uncertainty. Bishop argues this is a more *general* view of probabilities.

- Can be used when we are talking about a rare event. For example, the probability that the Arctic ice cap will have disappeared by the end of the century.

1.3.2 Bayes theorem

In the Bayesian approach, we can convert a prior probability into a posterior probability by incorporating the evidence provided by the observed data. This is especially applicable to scenarios in which we want to describe uncertainty of quantities such as the parameters $\boldsymbol{w}$, such as the polynomial curve-fitting example encountered earlier. We caputre our assumptions about $\boldsymbol{w}$, before observing the data, in the form of a prior distribution $p(\boldsymbol{w})$. The effect of observing the data $D = \{t_1, ..., t_n\}$ is expressed through the *likelihood* function $p(D|\boldsymbol{w})$. Bayes theorem,

$$p(\boldsymbol{w}|D) = \frac{p(D|\boldsymbol{w})p(\boldsymbol{w})}{p(D)} \tag{1}$$

then allows us to evaluate the uncertainty in $\boldsymbol{w}$ *after* we have observed D. This uncertainty in $\boldsymbol{w}$ is expressed through the *posterior* $p(\boldsymbol{w}|D)$.

- $p(D|\boldsymbol{w})$ is evaluated for the dataset D, and can be viewed as a function of the parameters $\boldsymbol{w}$. It represents how likely the data is, given different settings of $\boldsymbol{w}$.
- $p(D|\boldsymbol{w})$ is not a probability distribution over $\boldsymbol{w}$, and so $\int p(D|\boldsymbol{w})d\boldsymbol{w}$ does not necessarily equal 1.

1.3.3 Bayesian vs. Frequentist continued

The *likelihood* function $p(D|\boldsymbol{w})$ plays a central role in both Bayesian and Frequentist paradigms. But, how it is employed is fundamentally different...

- *Frequentist*: $\boldsymbol{w}$ is fixed and it is a *secret of nature* whose value is determined by the form of some "estimator", and error bars are created by considering the distribution of all possible D.
- *Bayesian*: D is fixed, and our uncertainty is expressed through a probability distribution over the parameters $\boldsymbol{w}$.

1.3.4 Frequentist examples of estimator and error bars

A widely used frequentist estimator is *maximum likelihood*, where $\boldsymbol{w}$ is set to the value that maximizes the likelihood $p(D|\boldsymbol{w})$. That is, it chooses $\boldsymbol{w}$ so that the probability of the observed data is maximized.

- In ML literature, the negative log of the likelihood is called an *error function*.

One approach to getting error bars is the *bootstrap*, where multiple datasets are created by drawing samples of D with replacement, and then looking at the variability of predictions between the different bootstrap datasets.

1.4 Maximum likelihood and the Gaussian [1.2.4]

1.4.1 Maximum likelihood in action

Suppose we have a set of observations $\mathbf{x} = \{x_1, ..., x_N\}$ constructed by drawing $i.i.d$ samples from a Gaussian distribution with unknown mean μ and variance σ^2. Our goal is to determine μ and σ^2 from the dataset.

Because are samples are $i.i.d$:

$$p(\mathbf{x}|\mu, \sigma^2) = \prod_{n=1}^{N} p(x_n|\mu, \sigma^2)$$

- When viewed as a function of μ, σ^2, this is the *likelihood* function for the Gaussian.

We will maximize the log likelihood, yielding the following maximum likelihood estimates:

$$\text{The sample mean} \qquad \mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n \tag{2}$$

$$\text{The sample variance} \qquad \sigma^2_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{ML})^2 \tag{3}$$

Later we will highlight the significant limitations of the maximum likelihood approach. However, here we will give a sneak peek of the issue, in the context of our solutions from above. Namely we will show,

- the maximum likelihood approach *systematically underestimates the variance of the distribution.*
- an example of a phenomenon called *bias* and is related to *overfitting.*

We first note that μ_{ML} and σ^2_{ML} are functions of the dataset $\{x_1, ..., x_N\}$. Consider the expectation of the quantities with respect to the dataset values, which themselves come from a Gaussian with parameters μ, σ^2.

$$\mathbb{E}[\mu_{ML}] = \mu \tag{4}$$

$$\mathbb{E}[\sigma^2_{ML}] = (\frac{N-1}{N})\sigma^2 \tag{5}$$

So on average the maximum likelihood approach will obtain the correct mean, but underestimate the true variance by a factor of $\frac{N-1}{N}$.

- This is because σ^2_{ML} is measured relative to the sample mean and not the true mean.

From (5), it follows that the following estimate for the variance is unbiased,

$$\tilde{\sigma}^2 = \frac{N}{N-1}\sigma^2_{ML} = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \mu_{ML})^2 \tag{6}$$

- The bias of maximum likelihood becomes less significant as the number of data points increases.
- For anything but small N, this bias won't be a serious issue, *however* later with more complex models that have many more parameters, the issue will be much more severe.
- In fact, we will see that the issue of bias in maximum likelihood lies at the root of the overfitting problem.

1.5 Curve fitting re-visited [1.2.5]

1.5.1 Maximum likelihood

Earlier we addressed the curve fitting problem by viewing it as an *error minimization problem*. We now look at it from a *probabilistic* perspective, where our final destination is the Full Bayesian treatment.

Recall the goal of the curve fitting problem: make predictions for a target t given some new value of input x on the basis of a set of N training data $\mathbf{x} = \{x_1, ..., x_N\}^T$ and the corresponding target values $\mathbf{t} = \{t_1, ..., t_N\}^T$.

We can express our uncertainty over the value of the *target variable* t using a probability distribution. That is, instead of having a point estimate for t, let our prediction be a distribution of possible t.

- To do this, we will assume that given the value of x, t follows a Gaussian with mean equal to $y(x, \boldsymbol{w})$ and variance β^{-1} (β is called the *precision*).
- Formally,

$$p(t|x, \boldsymbol{w}, \beta) = N(t|y(x, \boldsymbol{w}), \beta^{-1}) \tag{7}$$

- This is illustrated in Figure 1.

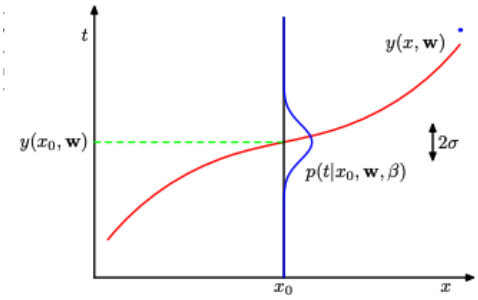

Figure 2: *Curve fitting problem assuming that $t|x = x_0$ is Gaussian with mean $y(x_0, \boldsymbol{w})$.*

We now use the training data $\{\mathbf{x}, \mathbf{t}\}$ to get the values of the unknown parameters β and $\boldsymbol{w}$ via maximum likelihood. Assuming that $\{\mathbf{x}, \mathbf{t}\}$ are *i.i.d*, the likelihood is then

$$p(\mathbf{t}|\mathbf{x}, \boldsymbol{w}, \beta) = \prod_{n=1}^{N} N(t_n|y(x_n, \boldsymbol{w}), \beta^{-1}) \tag{8}$$

As in before, we will use the (nicer) log likelihood function,

$$\ln p(\mathbf{t}|\mathbf{x}, \boldsymbol{w}, \beta) = -\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, \boldsymbol{w}) - t_n\}^2 + \frac{N}{2}\ln\beta - \frac{N}{2}ln(2\pi) \tag{9}$$

In getting $\boldsymbol{w}_{ML}$...

- Discard the last two terms and change $\frac{\beta}{2}$ to $\frac{1}{2}$.
- We also minimize the *negative* log likelihood instead of maximizing the log likelihood.
- Ending up with

$$\frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \boldsymbol{w}) - t_n\}^2 \tag{10}$$

  whic is the sum-of-squares error function.

In getting $\boldsymbol{w}_{ML}$, the sum-of-squares error function has arisen naturally as a consequence of maximizing the likelihood under the assumption of a Gaussian noise distribution over the target t conditioned on a given value of x.

We can then also determine β^{-1} with maximum likelihood by maximizing (9) with respect to β to yield

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

Now that we have $\boldsymbol{w}_{ML}$ and β_{ML}^{-1} we can make predictions for new x. Because we now have a probabilistic model, this is expressed as a *predictive distribution* that gives the probability distribution over t rather than a point estimate of t, and is obtained by subbing the maximum likelihood parameters into (7) to get

$$p(t|x, \boldsymbol{w}_{ML}, \beta_{ML}) = N(t|y(x, \boldsymbol{w}_{ML}), \beta_{ML}^{-1}) \tag{12}$$

Notice that we still have point estimates for $\boldsymbol{w}$ and we only consider the uncertainty in t that arises from unavoidable random noise in the labels. Therefore, we're only considering the aleatoric uncertainty in the above approach. Later, we will also incorporate our uncertainty in the parameters $\boldsymbol{w}$ when we reach the Full Bayesian treatment. On another note, (12) is the resultant *predictive distribution* from maximizing the likelihood generally denoted $p(D|\boldsymbol{w})$ but in the particular example above $p(D|\boldsymbol{w}) = p(\mathbf{t}|\mathbf{x}, \boldsymbol{w}, \beta)$ given by (8).

1.5.2 Maximum posteriori: MAP

We now take one step closer to the Fully Bayesian treatment by introducing a prior distribution over the weights $\boldsymbol{w}$. This prior denotes our beliefs about $\boldsymbol{w}$ *prior* to seeing the training data $\{\mathbf{x}, \mathbf{t}\}$. For simplicity, we consider a Gaussian prior of the form

$$p(\boldsymbol{w}|\alpha) = N(\boldsymbol{w}|0, \alpha^{-1}I) = (\frac{\alpha}{2\pi})^{(M+1)/2}exp\{-\frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w}\} \tag{13}$$

where α is the precision and $M+1$ is the total number of weights ($w \in \boldsymbol{w}$) for an M^{th} order polynomial. α is known as a hyper-parameter in this context.

From Bayes rule we know that the posterior over $\boldsymbol{w}$ is *proportional* to the *prior* $\times$ *likelihood*. In our example, it can be written as

$$p(\boldsymbol{w}|\boldsymbol{x}, \boldsymbol{t}, \alpha, \beta) \propto p(\boldsymbol{t}|\boldsymbol{w}, \boldsymbol{w}, \beta) \cdot p(\boldsymbol{w}|\alpha) \tag{14}$$

We can find the most *probable* $\boldsymbol{w}$ given the data by maximizing the right-hand-side of (14). This is known as *maximum posteriori* or *MAP*. In comparison to maximum likelihood, we've simply added prior beliefs about $\boldsymbol{w}$.

Taking the negative log of (14) and combining with the log likelihood in (9) and the prior in (13), we find that to get the *MAP* estimate, we minimize

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

So maximizing the posterior is equivalent to minimizing the *regularized* sum-of-squares error function, where the regularization parameter $\lambda=\frac{\alpha}{\beta}$.

Notice again that we still are getting a *point estimate* of $\boldsymbol{w}$.

1.5.3 Full Bayesian treatment

Though we've introduced a prior over the weights $p(\boldsymbol{w}|\alpha)$, we are still only getting a *point estimate* of $\boldsymbol{w}$ and so we are not yet Fully Bayesian! Recall that as Bayesians, we consider uncertainty to be in $\boldsymbol{w}$.

Here we arrive at our destination, the Fully Bayesian treatment. To do so, we should consistently apply the *sum* and *product* rules of probability which requires (as we shall see) that we integrate over all possible values of $\boldsymbol{w}$. This integration (marginalization) is at the heart of Bayesian methods for pattern recognition as it takes into account all possible $\boldsymbol{w}$ (one can think of this as taking into account all possible models) and their respective probabilities to address our uncertainty in our predictions later on.

For simplicity let us assume α and β are fixed and known. The Full Bayesian treatment then amounts to computing a predictive distribution $p(t|x,\mathbf{x},\mathbf{t})$ by consistent application of the *sum* and *product* rules so that we can write it as

$$p(t|x,\mathbf{x},\mathbf{t})=\int p(t|x,\boldsymbol{w})\cdot\underbrace{p(\boldsymbol{w}|\mathbf{x},\mathbf{t})}_{\text{posterior}}d\boldsymbol{w} \tag{16}$$

Here $p(t|x,\boldsymbol{w})$ is given by (7) and denotes the predictive distribution over t for an input x and a weight vector $\boldsymbol{w}$.

$p(\boldsymbol{w}|\mathbf{x},\mathbf{t})$ is the posterior, and incorporates our uncertainty over $\boldsymbol{w}$ given the data. We can view (16) as a distribution over t that takes into account all settings of $\boldsymbol{w}$ that are *possible* as computed in the posterior. Further, these settings are *weighted* by the posterior where more likely settings of $\boldsymbol{w}$ have a larger impact.

We shall see later that for problems such as the curve-fitting example, the posterior is Gaussian and can be evaluated analytically. Similarly, the integral above can be solved analytically resulting in a Gaussian predictive distribution.

Note that when the posterior is Gaussian, the mean corresponds to the mode and so the posterior mean and the MAP estimate for $\boldsymbol{w}$ is the same. Thus, the mean of their respective predictive distributions is the same. However, the variances of the their predictive distribution are different (Bayes considers both epistemic and aleatoric uncertainty while MAP only considers aleatoric). Additionally, there are many cases in which the posterior is not Gaussian and so predictions from *MAP* and the Fully Bayesian treatment will be different.

For our example, computing the right-hand-side of (16) leads to

$$p(t|x,\mathbf{x},\mathbf{t})=N(t|m(x),s^2(x)) \tag{17}$$

where the mean and variance are given by

$$m(x) = \beta \boldsymbol{\phi}(x)^T \boldsymbol{S} \sum_{n=1}^{N} \boldsymbol{\phi}(x_n) t_n \tag{18}$$

$$s^2(x) = \overbrace{\beta^{-1}}^{aleatoric} + \underbrace{\boldsymbol{\phi}(x)^T \boldsymbol{S} \boldsymbol{\phi}(x)}_{epistemic} \tag{19}$$

Here the matrix S is given by

$$S^{-1} = \alpha \boldsymbol{I} + \beta \sum_{n=1}^{N} \boldsymbol{\phi}(x_n) \boldsymbol{\phi}(x_n)^T \tag{20}$$

where I is the unit matrix and we have defined the vector $\boldsymbol{\phi}(x)$ with elements $\phi_i(x) = x^i$ for $i = 0, ..., M$.

We notice that as well as the mean, the variance of the predictive distribution in (17) depends on x. This is unlike maximum likelihood and *MAP*. The first term in (19) represents the uncertainty in t due to noise on the target variables; the second term represents the uncertainty in the parameters $\boldsymbol{w}$ and is a consequence of the Bayesian treatment.

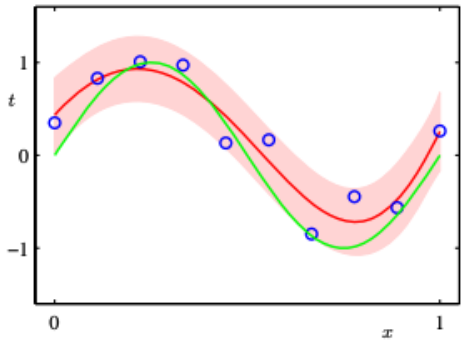

Figure 3: *Red line denotes the mean of the Bayesian predictive distribution given values of x. The red region denotes the ± 1 standard deviation around the mean. The variance varies depending on the value of x. The green line is the maximum likelihood predictive mean.*