ML, MAP, and Bayesian — The Holy Trinity of Parameter Estimation and Data Prediction

Avinash Kak
Purdue University

January 4, 2017
11:19am

An RVL Tutorial Presentation

originally presented in
Summer 2008

(minor changes in: January 2017)

Contents

Part 1: Introduction to ML, MAP, and Bayesian Estimation (Slides 3 – 28)

Part 2: ML, MAP, and Bayesian Prediction (Slides 29 – 33)

Part 3: Conjugate Priors (Slides 34 – 37)

Part 4: Multinomial Distributions (Slides 38 – 47)

Part 5: Modelling Text (Slides 49 – 60)

Part 6: What to Read Next? (Slides 61 – 62)

PART 1: Introduction to ML, MAP, and Bayesian Estimation

1.1: Say We are Given Evidence $\mathcal{X}$

Let's say that our evidence $\mathcal{X}$ consists of a set of **independent** observations:

$$\mathcal{X} \quad = \quad \Big\{\mathbf{x}_i\Big\}_{i=1}^{|\mathcal{X}|}$$

where each $\mathbf{x}_i$ is a realization of a random variable $\mathbf{x}$. [The notation $|\mathcal{X}|$ stands for the cardinality of $\mathcal{X}$, meaning the total number of observations in the set $\mathcal{X}$.] Each observation $\mathbf{x}_i$ is, in general, a data point in a multidimensional space.

Let's also say that a set Θ of probability distribution parameters best explains the evidence $\mathcal{X}$.

1.2: What Can We Do With The Evidence?

- We may wish to estimate the parameters Θ with the help of the Bayes' Rule

$$prob(\Theta|\mathcal{X}) \quad = \quad \frac{prob(\mathcal{X}|\Theta) \;\cdot\; prob(\Theta)}{prob(\mathcal{X})}$$

  where the notation $prob(A)$ stands for the probability of A and where $prob(A|B)$ means the conditional probability of A given B.

- Or, given a new observation $\tilde{\mathbf{x}}$, we may wish to compute the probability of the new observation being supported by the evidence:

$$prob(\tilde{\mathbf{x}}|\mathcal{X})$$

The former represents **parameter estimation** and the latter **data prediction**.

1.3: Focusing First on the Estimation of the Parameters Θ

We can interpret the Bayes' Rule

$$prob(\Theta|\mathcal{X}) \quad = \quad \frac{prob(\mathcal{X}|\Theta) \;\cdot\; prob(\Theta)}{prob(\mathcal{X})}$$

as

$$posterior \quad = \quad \frac{likelihood \;\cdot\; prior}{evidence}$$

Making explicit the formula for $likelihood$ as used above, we can write

$$likelihood \quad = \quad prob(\mathcal{X}|\Theta)$$

1.4: Maximum Likelihood (ML) Estimation of $\ominus$

We seek that value for $\ominus$ which maximizes the likelihood shown on the previous slide. That is, we seek that value for $\ominus$ which gives largest value to

$$prob(\mathcal{X}|\ominus)$$

We denote such a value of $\ominus$ by $\widehat{\ominus}_{ML}$.

We know that the joint probability of a collection of *independent* random variables is a product of the probabilities associated with the individual random variables in the collection.

Recognizing that the evidence $\mathcal{X}$ consists of the *independent* observations $\{\mathbf{x}_1, \mathbf{x}_2, ...\}$, we seek that value $\ominus$ which maximizes

$$\prod_{\mathbf{x}_i \in \mathcal{X}} prob(\mathbf{x}_i|\ominus)$$

Because of the product in the expression at the bottom of the previous slide, it is simpler to use its logarithm instead (since the logarithm is a monotonically increasing function of its argument).

Using the symbol $\mathcal{L}$ to denote the logarithm:

$$\mathcal{L} \quad = \quad \sum_{\mathbf{x}_i \in \mathcal{X}} log \ prob(\mathbf{x}_i | \Theta)$$

we can now write for the ML solution:

$$\widehat{\Theta}_{ML} \quad = \quad \underset{\Theta}{\text{argmax}} \ \mathcal{L}$$

That is, we seek those values for the parameters in Θ which maximize $\mathcal{L}$. The ML solution is usually obtained by setting

$$\frac{\partial \mathcal{L}}{\partial \theta_i} \quad = \quad 0 \qquad \forall \ \theta_i \in \Theta$$

1.5: Maximum a Posteriori (MAP) Estimation of Θ

For constructing the maximum *a posteriori* estimate for the parameter set Θ, we first go back to the Bayes' Rule on Slide 6:

$$prob(\Theta|\mathcal{X}) \quad = \quad \frac{prob(\mathcal{X}|\Theta) \;\cdot\; prob(\Theta)}{prob(\mathcal{X})}$$

We now seek that value for Θ which maximizes the posterior $prob(\Theta|\mathcal{X})$.

We denote such a value of Θ by $\widehat{\Theta}_{MAP}$.

Therefore, our solution can now be stated as shown on the next slide.

$$\begin{aligned}
\widehat{\Theta}_{MAP} &= \underset{\Theta}{\text{argmax}}\ prob(\Theta|\mathcal{X}) \\
&= \underset{\Theta}{\text{argmax}}\ \frac{prob(\mathcal{X}|\Theta)\ \cdot\ prob(\Theta)}{prob(\mathcal{X})} \\
&= \underset{\Theta}{\text{argmax}}\ prob(\mathcal{X}|\Theta)\ \cdot\ prob(\Theta) \\
&= \underset{\Theta}{\text{argmax}}\ \prod_{\mathbf{x}_i \in \mathcal{X}} prob(\mathbf{x}_i|\Theta)\ \cdot\ prob(\Theta)
\end{aligned}$$

As to why we dropped the denominator in the third re-write on the right, that's because it has no direct functional dependence on the parameters Θ with respect to which we want the right-hand side to be maximized.

As with the ML estimate, we can make this problem easier if we first take the logarithm of the posteriors. We can then write

$$\widehat{\Theta}_{MAP} = \underset{\Theta}{\text{argmax}} \left(\sum_{\mathbf{x}_i \in \mathcal{X}} log\ prob(\mathbf{x}_i|\Theta)\ +\ log\ prob(\Theta) \right)$$

1.6: What Does the MAP Estimate Get Us That the ML Estimate Does NOT

The MAP estimate allows us to inject into the estimation calculation our prior beliefs regarding the parameters values in Θ.

To illustrate the usefulness of such incorporation of prior beliefs, consider the following example provided by Gregor Heinrich:

Let's conduct N independent trials of the following Bernoulli experiment: *We will ask each person we see in the hallway outside this room whether they will vote Democratic or Republican in the next election.* Let p be the probability that an individual will vote Democratic.

In this example, each observation $\mathbf{x}_i$ is a scalar. So it's better to represent it by x_i. For each i, the value of x_i is either $Democratic$ or $Republican$.

We will now construct an ML estimate for the parameter p. The evidence $\mathcal{X}$ in this case consists of

$$\mathcal{X} = \left\{ x_i = \begin{array}{|l} Democratic \\ \\ Republican \end{array} , \quad i = 1...N \right\}$$

The log likelihood function in this case is

$$\begin{aligned} log\ prob(\mathcal{X}|p) &= \sum_{i=1}^{N} log\ prob(x_i|p) \\ &= \sum_{i} log\ prob(x_i = Demo) \\ &\quad + \sum_{i} log\ prob(x_i = Repub) \\ &= n_d \cdot log\ p + (N - n_d) \cdot log\ (1-p) \end{aligned}$$

where n_d is the number of individuals who are planning to vote Democratic this fall.

Setting

$$\mathcal{L} \quad = \quad log\ prob(\mathcal{X}|p)$$

we find the ML estimate for p by setting

$$\frac{\partial \mathcal{L}}{\partial p} \quad = \quad 0$$

That gives us the equation

$$\frac{n_d}{p} \ - \ \frac{(N - n_d)}{(1-p)} \quad = \quad 0$$

whose solution is the ML estimate

$$\widehat{p}_{ML} \quad = \quad \frac{n_d}{N}$$

So if $N = 20$ and if 12 out of 20 said that they were going to vote democratic, we get the following the ML estimate for p: $\widehat{p}_{ML} = 0.6$.

Now let's try to construct a MAP estimate for p for the same Bernoulli experiment.

Obviously, we now need a prior belief distribution for the parameter p to be estimated.

Our prior belief in possible values for p must reflect the following constraints:

- The prior for p must be zero outside the $[0, 1]$ interval.

- Within the $[0, 1]$ interval, we are free to specify our beliefs in any way we wish.

- In most cases, we would want to choose a distribution for the prior beliefs that peaks somewhere in the $[0, 1]$ interval.

The following **beta distribution** that is parameterized by two "shape" constants α and β does the job nicely for expressing our prior beliefs concerning p:

$$prob(p) = \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

where $B = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ is the beta function, with $\Gamma()$ denoting the Gamma function. The Gamma function $\Gamma()$ is a generalization of the notion of factorial to the case of real numbers. [The probability distribution shown above is also expressed as $Beta(p|\alpha,\beta)$.]

When both α and β are greater than zero, the above distribution has its mode — meaning its maximum value — at the following point

$$\frac{\alpha-1}{\alpha+\beta-2}$$

Let's now assume that we want the prior for p to reflect the following belief: *The state of Indiana (where Purdue is located) has traditionally voted Republican in presidential elections. However, on account of the prevailing economic conditions, the voters are more likely to vote Democratic in the election in question.*

We can represent the above belief by choosing a prior distribution for p that has a peak at 0.5. Setting $\alpha = \beta$ gives us a distribution for p that has a peak in the middle of the $[0, 1]$ interval.

As a further expression of our beliefs, let's now make the choice $\alpha = \beta = 5$. As to why, note that the variance of a beta distribution is given by

$$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

When $\alpha = \beta = 5$, we have a variance of roughly 0.025, implying a standard deviation of roughly 0.16, which should do for us nicely.

To construct a MAP estimate for p, we will now substitute the beta distribution prior for p in the following equation at the bottom of Slide 10:

$$\widehat{p}_{MAP} = \underset{p}{\text{argmax}} \left(\sum_{x \in \mathcal{X}} log\ prob(x|p) \ + \ log\ prob(p) \right)$$

which, with the help of the same rationale as used on Slide 12, can be rewritten for our specific experiment in the following form

$$\begin{aligned} \widehat{p}_{MAP} \ = \quad \underset{p}{\text{argmax}} \Big(& \ n_d \cdot log\ p \\ & + (N - n_d) \cdot log\ (1-p) \\ & + log\ prob(p) \ \Big) \end{aligned}$$

We can now substitute in the above equation the beta distribution for $prob(p)$ shown at the top of Slide 15. We must subsequently take the derivative of the right hand side of the equation with respect to the parameter p and set it to zero for finding best value for $\widehat{p}_{MAP}$.

The steps mentioned at the bottom of the previous slide give us the following equation:

$$\frac{n_d}{p} - \frac{(N - n_d)}{(1-p)} + \frac{\alpha - 1}{p} - \frac{\beta - 1}{1-p} = 0$$

The solution of this equation is

$$\widehat{p}_{MAP} = \frac{n_d + \alpha - 1}{N + \alpha + \beta - 2}$$

$$= \frac{n_d + 4}{N + 8}$$

With $N = 20$ and with 12 of the 20 saying they would vote Democratic, the MAP estimate for p is 0.571 with α and β both set to 5.

The next slide summarizes what we get from a MAP estimate beyond what's provided by an ML estimate.

- MAP estimation "pulls" the estimate toward the prior.

- The more focused our prior belief, the larger the pull toward the prior. By using larger values for α and β (but keeping them equal), we can narrow the peak of the beta distribution around the value of $p = 0.5$. This would cause the MAP estimate to move closer to the prior.

- In the expression we derived for $\widehat{p}_{MAP}$, the parameters α and β play a "**smoothing**" role vis-a-vis the measurement n_d.

- Since we referred to p as the *parameter* to be estimated, we can refer to α and β as the **hyperparameters** in the estimation calculations.

1.7: Bayesian Estimation

Given the evidence $\mathcal{X}$, ML considers the parameter vector Θ to be a constant and seeks out that value for the constant that provides maximum support for the evidence. ML does NOT allow us to inject our prior beliefs about the likely values for Θ in the estimation calculations.

MAP allows for the fact that the parameter vector Θ can take values from a distribution that expresses our prior beliefs regarding the parameters. MAP returns that value for Θ where the probability $prob(\Theta|\mathcal{X})$ is a maximum.

Both ML and MAP return only single and specific values for the parameter Θ.

Bayesian estimation, by contrast, calculates fully the posterior distribution $prob(\Theta|\mathcal{X})$.

Of all the Θ values made possible by the estimated posterior distribution, it is our job to select a value that we consider best in some sense. For example, we may choose the expected value of Θ assuming its variance is small enough.

The variance that we can calculate for the parameter Θ from its posterior distribution allows us to express our confidence in any specific value we may use as an estimate. If the variance is too large, we may declare that there does not exist a good estimate for Θ.

1.8: What Makes Bayesian Estimation Complicated?

Bayesian estimation is made complex by the fact that now the denominator in the Bayes' Rule

$$prob(\Theta|\mathcal{X}) \quad = \quad \frac{prob(\mathcal{X}|\Theta) \;\cdot\; prob(\Theta)}{prob(\mathcal{X})}$$

cannot be ignored. The denominator, known as the **probability of evidence**, is related to the other probabilities that make their appearance in the Bayes' Rule by

$$prob(\mathcal{X}) \;=\; \int_{\Theta} prob(\mathcal{X}|\Theta) \cdot prob(\Theta) \; d\Theta$$

This leads to the following thought critical to Bayesian estimation: **For a given likelihood function, if we have a choice regarding how we express our prior beliefs, we must use that form which allows us to carry out the integration shown at the bottom of the previous slide.** *It is this thought that leads to the notion of conjugate priors.*

Finally, note that, as with MAP, Bayesian estimation also requires us to express our prior beliefs in the possible values of the parameter vector Θ in the form of a distribution.

IMPORTANT PRACTICAL NOTE: Obtaining an algebraic expression for the posterior is, of course, important from a theoretical perspective. In practice, if you estimate a posterior ignoring the denominator, you can always find the normalization constant — which is the role served by the denomicator — simply by adding up what you get for the numerator, assuming you did a sufficiently good job of estimating the numerator. More on this point is in my tutorial "Monte Carlo Integration in Bayesian Estimation."

1.9: An Example of Bayesian Estimation

We will illustrate Bayesian estimation with the same Bernoulli trial based example we used earlier for ML and MAP. Our prior for that example is given by the following beta distribution:

$$prob(p|\alpha,\beta) \;\; = \;\; \frac{1}{B(\alpha,\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

where the LHS makes explicit the dependence of the prior on the hyperparameters α and β.

With this prior, the probability of evidence, defined on Slide 22, is given by

$$\begin{aligned} prob(\mathcal{X}) &= \int_0^1 prob(\mathcal{X}|p) \cdot prob(p)\ dp \\ &= \int_0^1 \left(\prod_{i=1}^{N} prob(x_i|p) \right) \cdot prob(p)\ dp \\ &= \int_0^1 \left(p^{n_d} \cdot (1-p)^{N-n_d} \right) \cdot prob(p)\ dp \end{aligned}$$

As it turns out, the integration at the bottom of the previous slide is easy.

When we multiply a beta distribution with either a power of p or a power of $(1-p)$, you simply get a different beta distribution.

So the probability of evidence for this example can be thought of as a constant Z whose value depends on the values chosen for α, β, and the measurement n_d.

We can now go back to the expression on the right side of the equation for Bayesian estimation, as shown by the first equation on Slide 22, and replace its denominator by Z as defined above.

With the step mentioned at the bottom of the previous slide, the Bayes' Rule for Bayesian estimation shown on Slide 22 becomes:

$$
\begin{aligned}
prob(p|\mathcal{X}) &= \frac{prob(\mathcal{X}|p) \;\cdot\; prob(p)}{Z} \\
&= \frac{1}{Z} \cdot prob(\mathcal{X}|p) \cdot prob(p) \\
&= \frac{1}{Z} \cdot \left(\prod_{i=1}^{N} prob(x_i|p) \right) \cdot prob(p) \\
&= \frac{1}{Z} \cdot \left(p^{n_d} \cdot (1-p)^{N-n_d} \right) \cdot prob(p) \\
&= Beta(p \mid \alpha + n_d, \beta + N - n_d)
\end{aligned}
$$

where the last result follows from the observation made earlier that a beta distribution multiplied by either a power of p or a power of $(1-p)$ remains a beta distribution, albeit with a different pair of hyperparameters. [Recall the definition of $Beta()$ on Slide 15.]

As shown on Slide 15, the notation $Beta()$ in the last equation of the previous slide is a short form for the same beta distribution you saw before. The hyperparameters of this beta distribution are shown to the right of the vertical bar in the argument list.

The result on the previous slide gives us a closed form expression for the posterior distribution for the parameter to be estimated.

If we wanted to return a single value as an estimate for p, that would be the expected value of the posterior distribution we just derived. Using the standard formula for the expectation of a beta distribution, the expectation is given by expression shown at the top of the next slide.

$$\begin{aligned} \widehat{p}_{Bayesian} &= E\{p|\mathcal{X}\} \\ &= \frac{\alpha + n_d}{\alpha + \beta + N} \\ &= \frac{5 + n_d}{10 + N} \end{aligned}$$

for the case when we set both α and β to 5.

When $N = 20$ and 12 out of 20 individuals report that they will vote Democratic, our Bayesian estimate yields a value of 0.567. Compare that to the MAP value of 0.571 and the ML value of 0.6.

One benefit of the Bayesian estimation is that we can also calculate the variance associated with the above estimate. One can again use the standard formula for the variance of a beta distribution to show that the variance associated with the Bayesian estimate is 0.0079.

PART 2: ML, MAP, and Bayesian Prediction

2.1: What is Prediction in the Context of ML, MAP, and Bayesian Estimation?

Let's say we are given the evidence

$$\mathcal{X} \quad = \quad \{\mathbf{x}_i\}_{i=1}^{|\mathcal{X}|}$$

and that next a new datum $\tilde{\mathbf{x}}$ comes along. We want to know as to what extent the new datum $\tilde{\mathbf{x}}$ is supported by the evidence $\mathcal{X}$.

To answer this question, we can try to calculate the probability

$$prob(\tilde{\mathbf{x}}|\mathcal{X})$$

and determine as to what extent the evidence $\mathcal{X}$ can **predict** the new datum $\tilde{\mathbf{x}}$. **Prediction is also referred to as regression.**

2.2: ML Prediction

We can write the following equation for the probabilistic support that the past data $\mathcal{X}$ provides to a new observation $\tilde{\mathbf{x}}$:

$$\begin{aligned} prob(\tilde{\mathbf{x}}|\mathcal{X}) &= \int_{\Theta} prob(\tilde{\mathbf{x}}|\Theta) \cdot prob(\Theta|\mathcal{X})\ d\Theta \\ &\approx \int_{\Theta} prob(\tilde{\mathbf{x}}|\widehat{\Theta}_{ML}) \cdot prob(\Theta|\mathcal{X})\ d\Theta \\ &= prob(\tilde{\mathbf{x}}|\widehat{\Theta}_{ML}) \end{aligned}$$

What this says is that the probability model for the new observation $\tilde{\mathbf{x}}$ is the same as for all previous observations that constitute the evidence $\mathcal{X}$. In this probability model, we set the parameters to $\widehat{\Theta}_{ML}$ to compute the support that the evidence lends to the new observation.

2.3: MAP Prediction

With MAP, the derivation on the previous slide becomes

$$
\begin{aligned}
prob(\tilde{\mathbf{x}}|\mathcal{X}) &= \int_{\Theta} prob(\tilde{\mathbf{x}}|\Theta) \cdot prob(\Theta|\mathcal{X})\ d\Theta \\
&\approx \int_{\Theta} prob(\tilde{\mathbf{x}}|\widehat{\Theta}_{MAP}) \cdot prob(\Theta|\mathcal{X})\ d\Theta \\
&= prob(\tilde{\mathbf{x}}|\widehat{\Theta}_{MAP})
\end{aligned}
$$

This is to be interpreted in the same manner as the ML prediction presented on the previous slide. The probabilistic support for the new data $\tilde{\mathbf{x}}$ is to be computed by using the same probability model as used for the evidence $\mathcal{X}$ but with the parameters set to Θ_{MAP}.

2.4: Bayesian Prediction

In order to compute the support $prob(\tilde{x}|\mathcal{X})$ that the evidence $\mathcal{X}$ lends to the new observation $\tilde{x}$, we again start with the relationship:

$$prob(\tilde{\mathbf{x}}|\mathcal{X}) = \int_{\Theta} prob(\tilde{\mathbf{x}}|\Theta) \cdot prob(\Theta|\mathcal{X}) \; d\Theta$$

but now we must use the Bayes' Rule for the posterior $prob(\Theta|\mathcal{X})$ to yield

$$prob(\tilde{\mathbf{x}}|\mathcal{X}) = \int_{\Theta} prob(\tilde{\mathbf{x}}|\Theta) \cdot \frac{prob(\mathcal{X}|\Theta) \cdot prob(\Theta)}{prob(\mathcal{X})} \; d\Theta$$

PART 3: Conjugate Priors

3.1: What is a Conjugate Prior?

As you saw, Bayesian estimation requires us to compute the full posterior distribution for the parameters of interest, as opposed to, say, just the value where the posterior acquires its maximum value. As shown already, the posterior is given by

$$prob(\Theta|\mathcal{X}) = \frac{prob(\mathcal{X}|\Theta) \cdot prob(\Theta)}{\int prob(\mathcal{X}|\Theta) \cdot prob(\Theta)\ d\Theta}$$

The most challenging part of the calculation here is the derivation of a closed form for the marginal in the denominator on the right.

For a given algebraic form for the likelihood, the different forms for the prior $prob(\Theta)$ pose different levels of difficulty for the determination of the marginal in the denominator and, therefore, for the determination of the posterior.

For a given likelihood function $prob(\mathcal{X}|\Theta)$, a prior $prob(\Theta)$ is called a **conjugate prior** if the posterior $prob(\Theta|\mathcal{X})$ has the same algebraic form as the prior.

Obviously, Bayesian estimation and prediction becomes much easier should the engineering assumptions allow a conjugate prior to be chosen for the applicable likelihood function.

When the likelihood can be assumed to be Gaussian, a Gaussian prior would constitute a conjugate prior because in this case the posterior would also be Gaussian.

You have already seen another example of a conjugate prior earlier in this review. For the Bernoulli trail based experiment we talked about earlier, the beta distribution constitutes a conjugate prior. As we saw there, the posterior was also a beta distribution (albeit with different hyperparameters).

As we will see later, when the likelihood is a multinomial, the conjugate prior is the Dirichlet distribution.

PART 4: Multinomial Distributions

4.1: When Are Multinomial Distributions Useful for Likelihoods?

Multinomial distributions are useful for modelling the evidence when each observation in the evidence can be characterized by count based features.

As a stepping stone to multinomial distributions, let's first talk about binomial distributions.

Binomial distributions answer the following question: Let's carry out N trials of a Bernoulli experiment with p as the probability of success at each trial. Let n be the random variable that denotes the number of times we achieve success in N trials. The questions is: What's the probability distribution for n?

The random variable n has binomial distribution that is given by

$$prob(n) \quad = \quad \binom{N}{n} p^n (1-p)^{N-n}$$

with the binomial coefficient $\binom{N}{n} = \frac{N!}{k!(N-n)!}$.

A multinomial distribution is a generalization of the binomial distribution.

Now instead of a binary outcome at each trial, we have k possible mutually exclusive outcomes at each trial. **Think of rolling a k-faced die (that is possibly biased)**.

At each trial, the k outcomes can occur with the probabilities p_1, p_2, ..., p_k, respectively, with the constraint that they must all add up to 1.

We still carry out N trials of the underlying experiment and at the end of the N trials we pose the following question: **What is the probability that we saw n_1 number of the first outcome, n_2 number of the second outcome, ..., and n_k number of the k^{th} outcome?** This probability is a multinomial distribution and is given by

$$prob(n_1, ..., n_k) = \frac{N!}{n_1!...n_k!} p_1^{n_1} \cdot \cdot p_k^{n_k}$$

with the stipulation that $\sum_{i=1}^{k} n_k = N$. The probability is zero when this condition is not satisfied. Note that there are only $k-1$ free variables in the argument to $prob()$ on the left hand side.

We can refer to the N trials we talked about above as constituting **one multinomial experiment** in which we roll the die N times as we keep count of the number of times we see the face with one dot, the number of times the face with two dots, the number of times the face with three dots, and so on.

If we wanted to actually measure the probability $prob(n_1, ..., n_k)$ experimentally, we would carry out a large number of multinomial experiments and record the number of experiments in which the first outcome occurs n_1 times, the second outcome n_2 times, and so on.

If we want to make explicit the conditioning variables in $prob(n_1, ..., n_k)$, we can express it as

$$prob(n_1, ..., n_k \mid p_1, p_2, ..., p_k, N)$$

This form is sometimes expressed more compactly as

$$Multi(\vec{n} \mid \vec{p}, N)$$

4.2: What is a Multinomial Random Variable?

A random variable W is a multinomial r.v. if its probability distribution is a multinomial. The vector $\vec{n}$ shown at the end of the last slide is a multinomial random variable.

A multinomial r.v. is a vector. Each element of the vector stands for a count for the occurrence of one of k possible outcomes in each trial of the underlying experiment.

Think of rolling a die 1000 times. At each roll, you will see one of six possible outcomes. In this case, $W = (n_1, ..., n_6)$ where n_i stands for the number of times you will see the face with i dots.

4.3: Multinomial Modelling of Likelihoods for Images and Text Data

If each image in a database can be characterized by the number of occurrences of a certain preselected set of features, then the database can be modeled by a multinomial distribution. Carrying out N trials of a k-outcome experiment would now correspond to examining N most significant features in each image and measuring the frequency of occurrence of each feature — assuming that the total number of distinct features is k. Therefore, each of the N features in each image must be one of the k distinct features. We can think of each image as the result of **one multinomial experiment**, meaning one run of N trials with each trial consisting of ascertaining the identities of the N most significant features in the image and counting the number of occurrences of the different features (under the assumption that there can only exist k different kinds of features).

A common way to characterize text documents is by the frequency of the words in the documents. Carrying out N trials of a k-outcome experiment could now correspond to recording the N most prominent words in a text file. If we assume that our vocabulary is limited to k words, for each document we would record the frequency of occurrence of each vocabulary word. We can think of each document as a result of **one multinomial experiment**.

For each of the above two cases, if for a given image or text file the first feature is observed n_1 times, the second feature n_2 times, etc., the likelihood probability to be associated with that image or text file would be

$$prob(image|p_1, ..., p_k) \quad = \quad \prod_{i=1}^{k} {p_i}^{n_i}$$

We will refer to this probability as the **multinomial likelihood** of the image. We can think of p_1, ... , p_k as the parameters that characterize the database.

4.4: Conjugate Prior for a Multinomial Likelihood

The conjugate **prior** for a multinomial likelihood is the Dirichlet distribution:

$$prob(p_1, ..., p_k) \quad = \quad \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} p_i^{\alpha_i - 1}$$

where α_i, $i = 1, ..., k$, are the hyperparameters of the prior.

The Dirichlet is a generalization of the beta distribution from two degrees of freedom to k degrees of freedom. (Strictly speaking, it is a generalization from the one degree of freedom of a beta distribution to $k-1$ degrees of freedom. That is because of the constraint $\sum_{i=1}^{k} p_i = 1$.)

The Dirichlet prior is also expressed more compactly as

$$prob(\vec{p}|\vec{\alpha})$$

For the purpose of visualization, consider the case when an image is only allowed to have three different kinds of features and we make N feature measurements in each image. In this case, $k = 3$. The Dirichlet prior would now take the form

$$prob(p_1, p_2, p_3 \mid \alpha_1, \alpha_2, \alpha_3)$$

$$= \frac{\Gamma(\alpha_1 + \alpha_2 + \alpha_3)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)} p_1^{\alpha_1 - 1} p_2^{\alpha_2 - 1} p_3^{\alpha_3 - 1}$$

under the constraint that $p_1 + p_2 + p_3 = 1$.

When $k = 2$, the Dirichlet prior reduces to

$$prob(p_1, p_2 \mid \alpha_1, \alpha_2)$$

$$= \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} p_1^{\alpha_1 - 1} p_2^{\alpha_2 - 1}$$

which is the same thing as the beta distribution shown earlier. Recall that now $p_1 + p_2 = 1$. Previously, we expressed the beta distribution as $prob(p|\alpha, \beta)$. The p there is the same thing as p_1 here.

PART 5: Modelling Text

5.1: A Unigram Model for Documents

Let's say we wish to draw N prominent words from a document for its representation.

We assume that we have a vocabulary of V prominent words. Also assume for the purpose of mental comfort that $V << N$. (This latter assumption is not necessary for the theory to work.)

For each document, we measure the number of occurrences n_i for $word_i$ of the vocabulary. It must obviously be the case the $\sum_{i=1}^{V} n_i = N$.

Let the multinomial r.v. W represent the word frequency vector in a document. So for a given document, the value taken by this r.v. can be shown as the vector $(n_1, ..., n_V)$.

Let our corpus (database of text documents) be characterized by the following set of probabilities: The probability that $word_i$ of the vocabulary will appear in any given document is p_i. We will use the vector $\vec{p}$ to represent the vector $(p_1, p_2, ..., p_V)$. The vector $\vec{p}$ is referred to as defining the **Unigram statistics** for the documents.

We can now associate the following multinomial likelihood with a document for which the r.v. W takes on the specific value $\mathcal{W} = (n_1, ..., n_V)$:

$$prob(\mathcal{W}|\vec{p}) \quad = \quad \prod_{i=1}^{V} p_i^{n_i}$$

If we want to carry out a Bayesian estimation of the parameters $\vec{p}$, it would be best if we could represent the priors by the Dirichlet distribution:

$$prob(\vec{p}|\vec{\alpha}) \quad = \quad Dir(\vec{p}|\vec{\alpha})$$

Since the Dirichlet is a conjugate prior for a multinomial likelihood, our posterior will also be a Dirichlet.

Let's say we wish to compute the posterior after observing a single document with $\mathcal{W} = (n_1, ..., n_V)$ as the value for the r.v. W. This can be shown to be

$$prob(\vec{p}|\mathcal{W}, \vec{\alpha}) \quad = \quad Dir(\vec{p}|\vec{\alpha} + \vec{n})$$

5.2: A Mixture of Unigrams Model for Documents

In this model, we assume that the word probability p_i for the occurrence of $word_i$ in a document is conditioned on the selection of a **topic** for a document. In other words, we first assume that a document contains words corresponding to a specific topic and that the word probabilities then depend on the topic.

Therefore, if we are given, say, 100,000 documents on, say, 20 topics, and if we can assume that each document pertains to only one topic, then the mixture of unigrams approach will partition the corpus into 20 clusters by assigning one of the 20 topics labels to each of the 100,000 documents.

We can now associate the following likelihood with a document for which the r.v. W takes on the specific value $\mathcal{W} = (n_1, ..., n_V)$, with n_i as the number of times $word_i$ appears in the document:

$$prob(\mathcal{W}|\vec{p}, \vec{z}) \;=\; \sum_{z} prob(z) \cdot \prod_{i=1}^{V} (p(word_i|z))^{n_i}$$

Note that the summation over the topic distribution does not imply that a document is allowed to contain multiple topics simultaneously. It simply implies that a document, constrained to contain only one topic, may either contain topic z_1, or topic z_2, or any of the other topics, each with a probability that is given by $prob(z)$.

5.3: Document Modelling with PLSA

PLSA stands for Probabilistic Latent Semantic Analysis.

PLSA extends the mixture of unigrams model by considering the document itself to be a random variable and declaring that a document and a word in the document *are conditionally independent if we know the topic that governs the production of that word.*

The mixture of unigrams model presented on the previous slide required that a document contain only one topic.

On the other hand, with PLSA, as you "generate" the words in a document, at each point you first randomly select a topic and then select a word based on the topic chosen.

The topics themselves are considered to be the hidden variables in the modelling process.

With PLSA, the probability that the word w_n from our vocabulary of V words will appear in a document d is given by

$$prob(d, w_n) \quad = \quad prob(d) \sum_z prob(w_n|z) prob(z|d)$$

where the random variable z represents the hidden topics. Now a document can have any number of topics in it.

5.4: Modelling Documents with LDA

LDA is a more modern approach to modelling documents.

LDA takes a more principled approach to expressing the dependencies between the topics and the documents on the one hand and between the topics and the words on the other.

LDA stands for **Latent Dirichlet Allocation**. The name is justified by the fact that the topics are the latent (hidden) variables and our document modelling process must allocate the words to the topics.

Assume for a moment that we know that a corpus is best modeled with the help of k topics.

For each document, LDA first "constructs" a multinomial whose k outcomes correspond to choosing each of the k topics. For each document we are interested in the frequency of occurrence of of each of the k topics. Given a document, the probabilities associated with each of the topics can be expressed as

$$\theta \quad = \quad [p(z_1|doc), ..., p(z_k|doc)]$$

where z_i stands for the i^{th} topic. It must obviously be the case that $\sum_{i=1}^{k} p(z_i|doc) = 1$. So the θ vector has only $k-1$ degrees of freedom.

LDA assumes that the multinomial θ can be given a Dirichlet prior:

$$prob(\theta|\alpha) \quad = \quad \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} \theta_i^{\alpha_i - 1}$$

where θ_i stands for $p(z_i|doc)$ and where α are the k hyperparameters of the prior.

Choosing θ for a document means randomly specifying the topic mixture for the document.

After we have chosen θ randomly for a document, we need to generate the words for the document. This we do by first randomly choosing a topic at each word position according to θ and then choosing a word by using the distribution specified by the β matrix:

$$\beta = \begin{bmatrix} p(word_1|z_1) & p(word_2|z_1) & ... & p(word_V|z_1) \\ p(word_1|z_2) & p(word_2|z_2) & ... & p(word_V|z_2) \\ ... & & ... & ... \\ ... & & ... & ... \\ p(word_1|z_K) & p(word_2|z_K) & ... & p(word_V|z_K) \end{bmatrix}$$

What is interesting is that these probabilities cut across all of the documents in the corpus. *That is, they characterize the entire corpus.*

Therefore, a corpus in LDA is characterized by the parameters α and β.

Folks who do research in LDA have developed different strategies for the estimation of these parameters.

PART 6: What to Read Next?

Recommendations for Further Reading

- If you are interested in recent research results on the precision with which bugs can be automatically localized (using bug reports as queries) through the modeling of large software libraries using the methods of Part 5 of this tutorial, see the slides of my recent talk on the subject: **"Importance of Machine Learning to the SCUM of Large Software"** at `https://engineering.purdue.edu/kak/AviKakInfyTalk2013_Handout.pdf`

- If you would like to go deeper into the practical aspects of Bayesian estimation, you might wish to read my tutorial **"Monte Carlo Integration in Bayesian Estimation,"** that is available at `https://engineering.purdue.edu/kak/Tutorials/MonteCarloInBayesian.pdf`

Acknowledgments

This presentation was a part of a tutorial marathon we ran in Purdue Robot Vision Lab in the summer of 2008. The marathon consisted of three presentations: my presentation that you are seeing here on the foundations of parameter estimation and prediction, Shivani Rao's presentation on the application of LDA to text analysis, and Gaurav Srivastava's presentation of the application of LDA to scene interpretation in computer vision.

This tutorial presentation was inspired by the wonderful paper "Parameter Estimation for Text Analysis" by Gregor Heinrich that was brought to my attention by Shivani Rao. My explanations and the examples I have used follow closely those that are in the paper by Gregor Heinrich.

This tutorial presentation has also benefitted considerably from my many conversations with Shivani Rao.