Contents

1 Basics

1.1 Random Variables

We will denote random variables with capital letters, X, Y, Z, and the set of values they can take as $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$. The specific value that variables X, Y, Z took on sample i will usually be denoted by lowercase letters x_i, y_i, z_i.

Unless explicitly noted we will want to leave $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ are very general. For example they could be

A binary choice $\{0, 1\}$ the set with two elements 0 and 1.

A finite set of categories $\{A, B, C, D, E, \dots\}$, no ordering or relationship of any kind is implied between the different categories.

the real numbers $\mathbb{R}$.

An open numeric interval i.e. $(0, 1)$ the numbers between 0 and 1 (excluding the end points).

real, ordered K-tuples $\mathbb{R}^K$, $(x_1, \dots, x_k, \dots, x_K)$, where $x_k \in \mathbb{R}$.

binary, ordered K-tuples $(x_1, \dots, x_k, \dots, x_K)$, where $x_k \in \{0, 1\}$.

A word in the English language

Any document downloaded from the internet

A color image with $P \times P$ resolution and colors described as RBG

In practice, thinking of $\mathcal{X}$ as been either a finite set of categories $A, B, \dots, K$, or $\mathcal{R}^K$ will provide most of the intuition.

1.2 Conditional Probability

The probability of that random variable X takes value x will be denoted as $P_X(x)$. If x is a continuous variable, this is to be interpreted as a probability density function. We will drop the subscript when the random variable is clear from the context.

Joint Probability $P(Y, X)$:

Probability that both X and Y happen at the same time.

This is a probability distribution over X and Y, normalized as

$$\int \mathrm{d}X \mathrm{d}Y \, P(Y, X) = 1 \tag{1}$$

Conditional Probability $P(Y|X)$

Probability that Y happen given that X has happened.

This is a family of probability distributions over Y (one for each possible value of X) normalized as

$$\int \mathrm{d}Y \, P(Y \mid X) = 1 \tag{2}$$

Marginal Probability P(Y)

Probability of Y ignoring the value of X. This is a single probability distribution

$$\int \mathrm{d}Y \, P(Y) = 1 \tag{3}$$

1.3 Bayes Theorem

Bayes theorem is central to statistical inference, it can be shown in may forms

$$\begin{aligned} P(Y,X) &= P(Y \mid X)P(X) \\ P(Y \mid X) &= \frac{P(Y,X)}{P(X)} \\ P(X \mid Y) &= \frac{P(Y,X)}{P(Y)} \\ P(Y \mid X) &= P(X \mid Y)\frac{P(Y)}{P(X)} \end{aligned} \tag{4}$$

1.4 Expectation, Variance

The (population) expected value of a function $f(x)$ is

$$\bar{f} = \mathbb{E}_X(f) = \int \mathrm{d}x P_X(x) f(x), \tag{5}$$

We will drop the subscript when there is no ambiguity.
Similarly the (population) variance is defined as

$$\mathrm{Var}(f) = \mathbb{E}_X((f-\bar{f})^2) = \int \mathrm{d}x P_X(x) \left(f(x)-\bar{f}\right)^2 \tag{6}$$

Given a set of independent, identically distributed random variables $X_1, X_2, \ldots, X_N$, and their observations $x_1, x_2, \ldots x_N$ we define the sample mean as

$$\hat{f} = \frac{1}{N}\sum_{i=1}^{N} f(x_i) \tag{7}$$

And the sample variance as

$$\widehat{Var} f = \frac{1}{N-1}\sum_{i=1}^{N}\left(f(x_i)-\hat{f}\right)^2 \tag{8}$$

1.5 Central Limit Theorem

Given **independent**, **identically distributed** variables $X_1, X_2, \ldots X_N$ with mean

$$\mathbb{E}(x_i) = \mu \tag{9}$$

and variance

$$\mathrm{Var}(x_i) = \sigma^2 \tag{10}$$

The sum

$$S_N = \sum_{i=1}^{N} X_i \tag{11}$$

has a Gaussian limit distribution

$$S_N \sim \mathcal{N}(N\mu, N\sigma^2) \tag{12}$$

The variables X_i do not need to be Gaussian, just have finite variance.
The sample mean is then

$$\hat{x} = \frac{1}{N} S_N = \frac{1}{N} \sum_{i=1}^{N} X_i \tag{13}$$

is also asymptotically Gaussian with distribution

$$\sqrt{N} \left(\frac{\hat{x} - \mu}{\sigma} \right) \sim \mathcal{N}(0, 1) \tag{14}$$

2 Statistical Inference

Let's assume we have a random variable $S \in \mathcal{S}$, and we have a family of probability distributions $P(s; \theta)$ parameterized by a set of parameters $\theta \in \Theta$

We have two kinds of inference problems we will interested on

Parameter Estimation of θ based on some, independently observed samples $\{x_i\}$

Predictive Posterior Distribution Estimation of a new, unobserved, s_{N+1} given previously observed samples s_i, for $i = 1, \ldots, N$

$$P(X_{N+1} \mid X_1, X_2, \ldots X_N) \tag{15}$$

2.1 Maximum Likelihood Inference

In maximum lilekihood estimation we assume samples $s_i = i, \ldots, N$ of random variable S are generated independently from $P(s; \theta_0)$ with a fixed (but unknown) value of the parameter θ_0.

2.1.1 Parameter Estimation

To estimate θ_0 we consider the **Likelihood** function. This is nothing but the probability of observing sample s_i, considered as a function of θ. Given the independence of the observations this can be written as

$$\text{Likelihood}(\theta; \{s_i\}_{i=1}^N) = \prod_{i=1}^{N} P(s_i; \theta) \tag{16}$$

The maximum likelihood estimate (MLE) of θ, which we will denote by $\hat{\theta}_{\text{MLE}}$ is the value of θ that maximizes the likelihood of the observed data s_i

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} \left[\text{Likelihood}(\theta, s_i) \right] \tag{17}$$

where we have introduced $\arg\max_\theta[f(\theta)]$ a function that returns the value θ where $f(\theta)$ achieves its maximum.

Dealing with the product on Equation 16 is cumbersome, so it is customary to work instead with the log of the likelihood function

$$\hat{l}(\theta; \{s_i\}) = \frac{1}{N}\sum_{i=1}^{N} \log P(s_i; \theta) = \frac{1}{N}\sum_{i=1}^{N} l(\theta; s_i) \tag{18}$$

where we have defined the log likelihood function over one observation as $l(\theta; s) = \log P(s; \theta)$.

As log is an increasing function and the normalization factor $1/N$ does not depend on θ we still have that

$$\hat{\theta}_{\text{MLE}} = \arg\max_\theta \left[\hat{l}(\theta, \{s_i\}) \right] \tag{19}$$

Asymptotic Properties of Maximum Likelihood the likelihood function Eq 18 is an average over the observations s_i (what we will call a sample average) of the function $l(s; \theta)$.

Because L is the sum of independent random variables, $\hat{l}$ is approximately normally distributed for large sample size N. And we have that

$$\lim_{N\to\infty} \hat{l}(\theta, \{s_i\}_{i=1}^N) = l(\theta) \tag{20}$$

the sample log likelihood function converges to the population log likelihood over the true distribution

$$l(\theta) = \mathbb{E}[\log P(s; \theta) \mid \theta_0] = \int \mathrm{d}s\, P(s; \theta_0) l(\theta; s) \tag{21}$$

When $\theta = (\theta_1, \theta_2, \dots \theta_K) \in \mathcal{R}^K$ and $l(\theta; s)$ is a smooth function of θ, $\hat{\theta}$ must satisfy the first order extremal conditions

$$\frac{\partial \hat{l}(\hat{\theta}; \{s_i\})}{\partial \theta_k} = 0 \tag{22}$$

It can be shown that $\hat{\theta}$, considered as a random variable (through is dependence on the sample data s_i) is asymptotically normally distributed.

$$\sqrt{N}(\hat{\theta}_{\text{MLE}} - \theta_0) \sim \mathcal{N}(0, I^{-1}) \tag{23}$$

where I is the Fisher information matrix

$$I_{j,k} = -\mathbb{E}\left[\frac{\partial^2 l}{\partial \theta_j \partial \theta_k} \mid \theta_0 \right] \tag{24}$$

Notation Remark In Machine learning we usually work with the *error function* $e = -l$. Everything is exactly the same, but we must minimize error instead of maximizing likelihood.

$$\hat{\theta} = \arg\min_\theta \frac{1}{N}\sum_{i=1}^{N} e(\theta; s_i) \tag{25}$$

2.1.2 Predictive Posterior Distribution Estimation

Once we have an estimate $\hat{\theta}$ our prediction for the probability of a new observation of S given the samples $\{s_i\}$ we have already observed is

$$P(s \mid \{s_i\}) = P(s; \hat{\theta}(\{s_i\}) \tag{26}$$

2.2 Bayesian Inference

In Bayesian inference instead of assuming there is one set of parameters θ_0 that generate our sample data $\{s_i\}$, we assume we have a *prior* probability distribution $P_0(\theta)$ representing our knowledge of the likely values of θ.

In this point of view θ is a random variable, and the sample points s_i are taken as a given.

2.2.1 Parameter Estimation

Using Bayes theorem to compute the *posterior* distribution of θ given the sample data

$$P(\theta \mid \{s_i\}) = \frac{P(\{s_i\} \mid \theta) P_0(\theta)}{p(s_i)} \tag{27}$$

where

$$p(s) = \int \mathrm{d}\theta P_0(\theta) P(s \mid \theta) \tag{28}$$

The following terminology is often used

Likelihood $P(\{s_i\} \mid \theta)$, this term the same terms that appear in maximum likelihood estimation.

Prior $P_0(\theta)$, measures our expectations about the parameter θ before we observe any sample data s_i.

Evidence $p(\{s_i\})$ is the marginal distribution of the observations $\{s_i\}$ given our prior.

From a Bayesian inference point of view, all the information about the problem is contained on the posterior distribution. The practical difficulty is that, except for the simplest cases the **evidence** term can not be computed analytically, and can be very expensive to approximate computationally.

Bayesian Point Estimates As the full posterior probability distribution $P(\theta|\{s_i\})$ may be hard to obtain, sometimes a single point estimate is generates as an *approximation*. There are a few choices possible, two of the more commont are

Maximum a posteriori (MAP)

$$\hat{\theta}_{\mathrm{MAP}} = \arg\max_{\theta} P(\theta|s_i) \tag{29}$$

As in the case of max likelihood this is more convenient to work with the log likelihood, using the fact that the evidence $p(\{s_i\})$ does not depend on θ, we can write

$$\hat{\theta}_{\text{MAP}} = \hat{l}(\theta; \{s_i\}) + \frac{1}{N}\Omega(\theta) \tag{30}$$

where $\Omega(\theta) = \log P_0(\theta)$ is usually called the penalty function.

Posterior Mean If $\theta \in \mathbb{R}^K$ so that it make sense to compute an expected value, we way use

$$\hat{\theta}_{\text{PM}} = \mathbb{E}(\theta \mid \{s_i\}) = \int \mathrm{d}\theta\, \theta P(\theta \mid \{s_i\}) \tag{31}$$

The θ_{MAP} has the practical advantage that one can reuse the same algorithms used for maximum likelihood optimization with just an extra $\Omega(\theta)$ penalty function on the objective function.

2.2.2 Predictive Posterior Distribution Estimation

One more use of Bayes theorem give us that

how do this come?

$$p(s \mid \{s_i\}) = \int \mathrm{d}\theta\, P(s \mid \theta) P(\theta \mid \{s_i\}) \tag{32}$$

If we replace the full posterior probability distribution by a point estimate we get again procedure equivalent to Maximum Likelihood

$$p(s \mid \{s_i\}) \approx P(s|\hat{\theta}_{\text{MAP}}(\{s_i\})) \tag{33}$$

or

$$p(s \mid \{s_i\}) \approx P(s|\hat{\theta}_{\text{PM}}(\{s_i\})) \tag{34}$$

but, in the Bayes inference context, this are just an approximations to the true answer Eq 32.

2.2.3 Conjugate Priors

In Bayesian inference our prior believes about θ are encoded into the prior distribution $P_0(\theta)$. It is usually convenient to select P_0 from a family of probability distributions parameterizer by yet another parameter that encodes our degree on uncertainty about the range of values that θ could have

$$P_0(\theta) = P_{\text{C}}(\theta; \alpha). \tag{35}$$

In this Bayesian setting, the parameter θ is a random variable that we will learn from the data, while the **hyper-parameter** α is considered fixed and determined exclusively by our previous beliefs.

In principle, any family of probability distributions would do, but, for *analytical* convenience, it is customary to select a family of such that, after conditioning on the observed data $\{s_i\}$, the posterior distribution still belongs to the same distribution family as the prior

$$P(\theta \mid \{\ s_i\}) = P_{\mathrm{C}}\left(\theta; \hat{\alpha}(\{s_i\}, \alpha_0)\right) = \frac{P(\{s_i\} \mid \theta) P_C(\theta, \mid \alpha_0)}{P(\{s_i\})} \tag{36}$$

Therefore, a Family of probability distributions satisfying

$$P_{\mathrm{C}}\left(\theta; \hat{\alpha}(\{s_i\}, \alpha_0)\right) = \frac{P(\{s_i\} \mid \theta) P_C(\theta, \mid \alpha_0)}{P(\{s_i\})} \tag{37}$$

is called a **conjugate prior** to the family of distributions $P(s \mid \theta)$.

In the next section we will show that the Beta distribution is the conjugate prior to the binomial distribution, and that the Gaussian distribution is conjugated to itself.

It is important to stress that conjugate priors are chosen for *analytic convenience* . They allow us to replace complicated equations containing integrals (Eq 27, Eq 28) with a much simpler parameter update as in Eq 37.

2.3 Supervised Learning as Statistical Inference

Let's assume that we have access to samples $\{y_i, x_i\}_{i=1}^N$ drawn from a tuple of random variables (Y, X), $Y \in \mathcal{Y}$, and $X \in \mathcal{X}$. As usual, we assume that samples for $i \neq j$ are independent of each other. Supervised learning's goal is to provide predictions for new (unseen) values of the target y_t given the sample data x_t.

In probabilistic terms, we want to learn $P(Y \mid X)$ given the sample data $\{y_i, x_i\}$. This is equivalent to infering $P(s)$ given sample data $\{s_i\}$ as we discussed in (Sec 2) but with the added difficulty that we need to learn a different probability function for each value of $X \in \mathcal{X}$. When the space $\mathcal{X}$ is large this can become quite challenging. We will spend the rest of the course working on this problem.

Maximum Likelihood estimation proceeds exactly as in section (2.1), with $P(s; \theta)$ replaced by $P(y \mid x; \theta)$. The conditional log likelihood function is $l(\theta; y, x) = \log P(y \mid x; \theta)$.

The maximum likelihood estimate for θ becomes

$$\hat{\theta} = \arg\max_{\theta} \hat{l}(\theta) = \arg\max_{\theta} \left[\frac{1}{N} \sum_{i=1}^{N} l(\theta; y_i, x_i) \right] \tag{38}$$

Bayesian estimation is exactly as in section (2.2) with $P(Y \mid X, \theta)$ replacing $P(S \mid \theta)$

3 Statistical Inference of Classical Distributions

3.1 Bernoulli Distribution

The simplest random variable takes values in the set binary set $\mathcal{S} = \{0, 1\}$ (for example the result of a a single coin toss). Its probability distribution is called the Bernoulli distribution and is completely characterized by a single parameter θ so that

$$\begin{aligned} P(s = 1; \theta) =& \theta \\ P(s = 0; \theta) =& 1 - \theta \end{aligned} \tag{39}$$

We collect this two expression into a single formula

$$P(s; \theta) = \theta^s (1 - \theta)^{1-s} \tag{40}$$

3.1.1 Maximum Likelihood Estimate

The likelihood function, thus is

$$l(\theta; s) = s \log \frac{\theta}{1 - \theta} + \log(1 - \theta) \tag{41}$$

and

$$\hat{l}(\theta) = \frac{1}{N} \sum_{i=1}^{N} s_i \log \frac{\theta}{1 - \theta} + \log(1 - \theta) = \frac{\hat{N}_1}{N} \log \frac{\theta}{1 - \theta} + \log(1 - \theta) \tag{42}$$

where $\hat{N}_1$ is the number of observations (also known as the count) where $s = 1$.

The MLE of θ must satisfy the first order condition

$$\frac{\partial \hat{l}}{\partial \theta} = \frac{\hat{N}_1}{N} \left(\frac{1}{\theta} - \frac{1}{1 - \theta} \right) - \frac{1}{1 - \theta} = 0 \tag{43}$$

solving for θ we obtain

$$\hat{\theta}_{\text{MLE}} = \frac{\hat{N}_1}{N} = \frac{\sum_{i=1}^{N} s_i}{N} \tag{44}$$

3.1.2 Bayes Estimation with Uninformative Prior

As the samples are independent the posterior distribution is

$$P(\theta | \{s_i\}) = \frac{\prod_{i=1}^{N} P(s_i \mid \theta)}{P(\{s_i\})} P_0(\theta). \tag{45}$$

If we assume a *non-informative* prior

$$P_0(\theta) = 1 \tag{46}$$

for $0 < \theta < 1$ so that all values of the probability are equally likely we have

$$P(\theta|\{s_i\}) = C\theta^{\hat{N}_1}(1-\theta)^{\hat{N}_0}, \tag{47}$$

where $\hat{N}_0 = N - \hat{N}_1$ is the number of samples where $s_i = 0$ and C is a normalization constant so that probabilities integrate to 1. ? ?

$$\frac{1}{C} = \int \mathrm{d}\theta\, \theta^{\hat{N}_1}(1-\theta)^{\hat{N}_0} = B(\hat{N}_1 + 1, \hat{N}_0 + 1), \tag{48}$$

where $B(\alpha, \beta)$ is the beta function.

The posterior distribution, thus, depends on the sample data only though the counts $\hat{N}_0$ and $\hat{N}_1$. In that situation we say that they are **sufficient statistic** for the binomial distribution. Given a sample $\{s_i\}$ of independent, identically distributed binomial variables, we can summarize it into just $\hat{N}_0$ and $\hat{N}_1$ without any loss of information.

MAP Estimate is the same as θ_{MLE} because the prior $P_0(\theta) = 1$ does not depend on θ

$$\theta_{\mathrm{MAP}} = \theta_{\mathrm{MLE}} = \frac{\hat{N}_1}{N} \tag{49}$$

Posterior Mean Estimate can be computed as

$$\theta_{\mathrm{PM}} = C\int \mathrm{d}\theta\, \theta P(\theta|\{s_i\}) = \int \mathrm{d}\theta^{\hat{N}_1+1}(1-\theta)^{\hat{N}_0} \tag{50}$$

which, is, again, a beta integral. Substituting C we find

$$\theta_{PM} = \frac{B(\hat{N}_1 + 2, \hat{N}_0 + 1)}{B(\hat{N}_1 + 1, \hat{N}_0 + 1)}, \tag{51}$$

using that

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)} \tag{52}$$

and that

$$\Gamma(\alpha + 1) = \alpha\Gamma(\alpha) \tag{53}$$

the expression for θ_{PM} can be simplified into

$$\theta_{\mathrm{PM}} = \frac{\hat{N}_1 + 1}{\hat{N}_0 + \hat{N}_1 + 2} \tag{54}$$

Predictive Posterior Distribution For the binomial distribution, the expected value of a new observation, given the current observations is

$$P(s = 1 \mid \{s_i\}) = \int \mathrm{d}\theta\, \theta P(\theta \mid \{s_i\}) = \hat{\theta}_{PM} = \frac{\hat{N}_1 + 1}{\hat{N}_0 + \hat{N}_1 + 2} \tag{55}$$

so, in this case, we can simply use θ_{PM} as out predictive probability.

We can see the Bayesian prediction as identical to the Maximum Likelyhood one if we define

$$\begin{aligned}\tilde{N}_0 &= \hat{N}_0 + 1 \\ \tilde{N}_1 &= \hat{N}_1 + 1\end{aligned} \tag{56}$$

where the extra 1 added to the possitive and negative case are called **pseudo-counts**. They are the result of our (diffuse) prior expectation for the value of θ.

When the number of observations $N = \hat{N}_0 + \hat{N}_1$ is large, the pseudo-counts become relatively small compare to the total evidence strenght N and MLE and Bayesian prediction converge to each other.

$$\theta_{\mathrm{PM}} = \theta_{\mathrm{MLE}} + O\left(\frac{1}{N}\right) \tag{57}$$

When the number of observations is small, however, the impact of the prior can be large.

The argument that, due to our diffuse prior, our expectation for the next observation of s should include the pseudo-counts

$$P(s = 1 \mid \{s_i\}) = \frac{\hat{N}_1 + 1}{\hat{N}_1 + \hat{N}_0 + 2} \tag{58}$$

is due to Laplace. As such, is referred to as **add-one Laplace Smoothing**.

3.1.3 Bayes Estimation with Beta Conjugate Prior

For the binomial probability distribution

$$P(s \mid \theta) = \theta^s (1-\theta)^{1-s} \tag{59}$$

The conjugate prior is the Beta distribution

$$P_\beta(\theta; \alpha, \beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}. \tag{60}$$

The uninformative prior used on the previous section 3.1.2 is a special case where $\alpha = 1$ and $\beta = 1$.

Given a set of observations $\{s_i\}$ with sufficient statistics $\hat{N}_1 = \sum_{i=1}^{N} s_i$, and $\hat{N}_0 = N - \hat{N}_1$, and using Eq 45 we have that

$$P(\theta \mid \{s_i\}) = C\theta^{\hat{N}_1+\alpha-1}(1-\theta)^{\hat{N}_0+\beta-1} \tag{61}$$

which is a $B(\alpha+\hat{N}_1, \beta+\hat{N}_0)$ distribution. Repeating the arguments in Sec 3.1.2 the predictive posterior is

$$P(s = 1 \mid \{s_i\}; \alpha, \beta) = \frac{\hat{N}_1 + \alpha}{\hat{N}_1 + \hat{N}_0 + \alpha + \beta} \tag{62}$$

The hyper-parameters α and β play the role of pseudo-counts: $\alpha+\beta$ is a measure of the strengh of our prior convictions measured as a number of observations; $\alpha/(\alpha+\beta)$ is our prior expectation for the value of θ.

In practice, common used hyper-parameter values are

- $\alpha = \beta = 1$: add-one Laplace smoothing. Every value of θ is equally likely a priory.
- $\alpha = \beta = 0.5$: add-half Laplace smoothing. The strength our our convictions is somewhat weaker and the prior is more concentrated on the extremes $\theta = 0$ and $\theta = 1$.

3.2 Multinomial Distribution

The simplest possible random variable is the Bernoulli distribution 3.1 able to take only values $\{0, 1\}$. The next level of generality is a random variable S able to take $k = 1, \cdots K$ different categorical values, with no implicit relationship between them.

The probability distribution of such a variable is characterized by K parameters θ_k such that

$$P(s = k \mid \theta) = \theta_k \tag{63}$$

where, to ensure normalization of the probability distribution we have the constrains

$$\sum_{k=1}^{K} \theta_k = 1, \quad 0 \leq \theta_k \tag{64}$$

A natural representation of the variable S is the **one-hot encoding** as a vector or random variables $Z = (Z_1, \ldots, Z_K)$ as

$$Z_k(S) = \mathbb{1}(S = k) \tag{65}$$

so that z_k is one if $s = k$ and z_k is zero other wise. The variables z_k are sometimes referred to as **dummy** variables.

With this representation we can re-write Eq 63 as

$$P(z \mid \theta) = \prod_{k=1}^{K} \theta_k^{z_k} \tag{66}$$

which is the generalization of Eq 40 from 2 to K possible values.

3.2.1 Maximum Likelihood Estimate

Generalizing the arguments in section 3.1.1 the likelihood function is

$$l(\theta; z) = \sum_{k=1}^{K} z_k \log \theta_k, \tag{67}$$

given a set of observations $\{s_i\}_{i=1}^N$ we have a matrix of dummies $z_{i,k}$, $i = 1, \dots N$, $k = 1, \dots K$ and a sample likelihood function 18

$$\hat{l}(\theta) = \frac{1}{N} \sum_{i=1}^{N} z_{i,k} \log \theta_k = \sum_k \frac{\hat{N}_k}{N} \log \theta_k \tag{68}$$

where the K sufficient statistics $\hat{N}_k$ are the number of observations with $z_{i,k} = 1$, i.e. $s_i = k$

$$\hat{N}_k = \sum_{i=1}^{N} z_{i,k} \tag{69}$$

The MLE of θ must minimize $\hat{l}(\theta)$ subject to the normalization constrain 64. Introducing Laplace's multiplier λ we have that

$$\mathcal{L}(\theta, \lambda) = \hat{l}(\theta) - \lambda \left(\sum_{k=1}^{K} \theta_k - 1 \right) \tag{70}$$

satisfies the first order conditions

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \lambda} &= 0 = \sum_k \theta_k - 1 \\ \frac{\partial \mathcal{L}}{\partial \theta_k} &= 0 = \frac{\hat{N}_k}{N\theta_k} - \lambda, \end{aligned} \tag{71}$$

a bit of algebra shows that

$$\hat{\theta}_k = \frac{\hat{N}_k}{N} \tag{72}$$

are the solutions.

In vector form this can be written as

$$\hat{\theta}_{\text{MLE}} = \frac{\hat{N}}{N} \tag{73}$$

which is again a generalization of Eq 44 from 2 to K cases.

3.2.2 Bayesian Estimation with Dirichlet Prior

The generalization of the Beta distribution to the multinomial setting is the Dirichlet distribution

$$P_{\text{D}}(\theta; \alpha) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} \theta^{\alpha_k - 1} \tag{74}$$

where the hyper-parameters α_k for $k = 1, \dots K$ play the role of pseudo-counts and $B(\alpha)$ is the multivariate generalization of the Beta function

$$B(\alpha) = \frac{\prod_k \Gamma(\alpha_k)}{\Gamma(\sum_k \alpha_k)}. \tag{75}$$

In term of the counts $\hat{N}_k$ the posterior distribution of θ is given by

$$P(\theta|\{s_i\};\alpha) = P_{\mathrm{D}}(\hat{N}_1 + \alpha_1, \ldots, \hat{N}_K + \alpha_K) \tag{76}$$

The predictive distribution is

$$P(z_k \mid \{s_i\};\alpha) = \frac{N_k + \alpha_k}{\sum_{j=1}^{K}(\hat{N}_j + \alpha_j)} \tag{77}$$

So, given the Dirichlet prior, Bayesian inference gives rise to a multivariate version of Laplace smoothing. Common choices are:

- Uninformative prior: $\alpha_k = 1$, and smothed prediction is

$$P(z_k \mid \{s_i\}) = \frac{\hat{N}_k + 1}{N + K}. \tag{78}$$

  the add-one Laplace smoothing, generalized to K classes.

- Central Dirichlet prior: all cases are, a priori, equally probable, therefore $\alpha_k = \alpha$. The strength of the prior is $K\alpha$ so that the smoothed predictive distribution becomes

$$P(z_k \mid \{s_i\};\alpha) = \frac{\hat{N}_k + \alpha}{N + \alpha K} \tag{79}$$

3.3 Univariate Gaussian Distribution

It will be instructive to first analyze inference for the univariate Gaussian before we move into the more general multivariate setting. In this, and the next subsection we will follow the exposition in [1].

Let's assume $S \in \mathbb{R}$ is distributed normally

$$P(s;\mu,\sigma^2) = \mathcal{N}(s;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(s-\mu)^2}{2\sigma^2}} \tag{80}$$

3.3.1 Maximum Likelihood Inference

In the notation of (Sec 2) $\theta = (\mu, \Sigma)$. The log likelihood function is

$$l_g(s;\mu,\sigma) = -\frac{(s-\mu)^2}{2\sigma^{-2}} - \frac{1}{2}\log 2\pi\sigma^2. \tag{81}$$

and the sample likelihood is

$$\hat{l}_g(\mu,\sigma^2;\{s_i\}) = \frac{1}{2N\sigma^2}\sum_{i=1}^{N}(s-\mu)^2 - \frac{1}{2}\log 2\pi\sigma^2. \tag{82}$$

Solving the first order conditions $\frac{\partial \hat{l}_g}{\partial \mu} = 0$ and $\frac{\partial \hat{l}_g}{\partial \sigma^2} = 0$ implies the well known results

$$\hat{\mu}_{\text{MLE}} = \frac{1}{N}\sum_i s_i$$
$$\hat{\sigma^2}_{\text{MLE}} = \frac{1}{N}\sum_i (s_i - \hat{\mu}_{\text{MLE}})^2 \tag{83}$$

3.3.2 Bayesian Inference

The Gaussian distribution has two parameters μ and σ^2 that play different roles and have different conjugate prior distributions. For the mean, the conjugate distribution is a Gaussian

$$P_\mu(\mu; \mu_0, \sigma_0^2) = \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}} \tag{84}$$

while, for σ^2, it is more convenient to think of the *precision* $\tau = \sigma^{-2}$, as the parameter. The precision τ has a natural gamma conjugate prior

$$P_\tau(\tau; \alpha, \beta) = \frac{\beta^\alpha \tau^{\alpha-1} e^{-\beta\tau}}{\Gamma(\alpha)}; \tag{85}$$

When doing Bayesian inference on the Gaussian distribution we can distinguish two cases:

Known Variance In this case, the conjugate distribution to the unknown mean is a Gaussian distribution

$$P_\mu(\mu; \mu_0, \sigma_0^2) = \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}} \tag{86}$$

and the posterior distribution after observations $\{s_i\}$ is

$$P(\mu \mid \{s_i\}, \mu_0, \sigma_0^2) = Ce^{-\frac{1}{2}\left(\frac{\sum_i (s_i-\mu)^2}{\sigma^2} + \frac{(\mu-\mu_0)^2}{\sigma_0^2}\right)} \tag{87}$$

Completing the square we can see that the Bayesian MAP and Posterior Mean estimates are equal and given by

$$\hat{\mu}_{PM} = \sigma_N^2 \left(\frac{\mu_0}{\sigma_0^2} + \frac{\sum_i s_i}{\sigma^2}\right) \tag{88}$$

where

$$\sigma_N^2 = \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}}. \tag{89}$$

The posterior distribution for μ is

$$P(\mu|\{s_i\}; \mu_0, \sigma_0^2) = \mathcal{N}(\mu; \mu_{\text{PM}}, \sigma_N^2), \tag{90}$$

and the predictive posterior is

$$P(s \mid \{s_i\}; \mu_0, \sigma_0^2) = \mathcal{N}(s; \mu_{\text{PM}}, \sigma_N^2 + \sigma^2) \tag{91}$$

Unknown Variance We assume that

$$P(s \mid \mu, \tau) = \mathcal{N}(s; \mu, \tau^{-1}) \tag{92}$$

where the parameters μ, τ follow the join **Normal-Gamma** distribution

$$P_{\mathrm{NG}}(\mu, \tau; \mu_0, \lambda_0, \alpha_0, \beta_0) = \frac{\beta_0^{\alpha_0}\sqrt{\lambda_0}}{\Gamma(\alpha_0)\sqrt{2\pi}} \tau^{\alpha_0 - \frac{1}{2}} e^{-\beta_0 \tau - \frac{\lambda_0 \tau (\mu - \mu_0)^2}{2}} \tag{93}$$

With a bit of algebra, it can be shown that the posterior probability for μ, τ is also a Normal-Gamma distribution with

$$P(\mu, \tau \mid \{s_i\}) = P_{\mathrm{NG}}(\mu, \tau; \hat{\mu}, \hat{\lambda}, \hat{\alpha}, \hat{\beta}) \tag{94}$$

where

$$\begin{aligned}
\hat{\mu} &= \frac{\lambda_0 \mu_0 + N \hat{\mu}_{\mathrm{MLE}}}{\lambda_0 + N} \\
\hat{\lambda} &= \lambda_0 + N \\
\hat{\alpha} &= \alpha_0 + \frac{N}{2} \\
\hat{\beta} &= \beta_0 + \frac{1}{2}\left(N\sigma^2_{\mathrm{MLE}} + \frac{\lambda_0 N (\hat{\mu}_{\mathrm{MLE}} - \mu_0)^2}{\lambda_0 + N} \right)
\end{aligned} \tag{95}$$

The posterior predictive is given by

$$P(s \mid \{s_i\}; \mu_0, \gamma_0, \alpha_0, \beta_0) = t_{2\hat{\alpha}}\left(s; \hat{\mu}, \frac{\hat{\beta}(\hat{\lambda} + 1)}{\hat{\alpha}\hat{\lambda}} \right) \tag{96}$$

Where t is the centered t-Student distribution

$$t_\nu(s; \mu, \sigma^2) = \frac{1}{\nu B(\frac{1}{2}, \frac{\nu}{2})} \left(1 + \frac{(s - \mu)^2}{\nu \sigma^2} \right)^{-\frac{\nu+1}{2}} \tag{97}$$

3.4 Multivariate Gaussian Distribution

The density of a multivariate Gaussian distribution with mean μ_d, $d = 1, \ldots D$ and $D \times D$ covariance matrix $\Sigma_{d,d'}$ is

$$\mathcal{N}(s; \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(s-\mu)^T \Sigma^{-1} (s-\mu)} \tag{98}$$

In the notation of (Sec 2) $\theta = (\mu, \Sigma)$. The log likelihood function is

$$l_G(s; \mu, \Sigma) = -\frac{1}{2}(s - \mu)^T \Sigma^{-1} (s - \mu) - \frac{1}{2} \log |\Sigma| - \frac{D}{2} \log(2\pi). \tag{99}$$

Solving for the first order extremal conditions $\frac{l}{\theta} = 0$ we find the MLE's

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} s_i$$
$$\hat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}(s-\mu)(s-\mu)^T \tag{100}$$

where we are using vector notation s_i and μ are a D-dimensional vectors.

Writing explicitly all the indexes $s_i = (s_{i,1}, s_{i,2}, \dots s_{i,D})$, etc the formulas become

$$\hat{\mu}_d = \frac{1}{N}\sum_{i=1}^{N} s_{i,d}$$
$$\hat{\Sigma}_{d,d'} = \frac{1}{N}\sum_{i=1}^{N}(s_{i,d}-\mu_d)(s_{i,d'}-\mu_{d'}) \tag{101}$$

Bayesian estimates are analogous to the univarate gaussian case **??**, we refer the reader to Chapter 4, pag 129 of [2] for the details.

4 χ^2 test for categorical data

Given a sample of categorical data $\{s_i\}_{i=1}^N$, with $s_i \in \mathcal{S} = \{L_1, L_2, \dots L_K\}$ different category labels we would like to test if the sample was generated from distribution

$$P_0(S = L_k) = p_k. \tag{102}$$

The number of observations of label k

$$\hat{O}_k = \sum_{i=1}^{N} \mathbb{1}(s_i = L_k) \tag{103}$$

is the sum on N multinomially distributed random variables The expected value of O_k is

$$E_k = Np_k \tag{104}$$

the following quantity

$$C^2 = \sum_{k=1}^{N} \frac{(\hat{O}_k - E_k)^2}{E_k} \tag{105}$$

is distributed as $C^2 \sim \chi^2_{K-1}$ random variable.

See [3] for a proof.

4.1 χ^2 test of independence for categorical variables

Let's assume categorical variables X with categories $A_1, A_2, \dots A_D$, and Y with categories $B_1, B_2, \dots B_K$ with distributions unknown.

We can use a variation of the χ^2 test above to test is X and Y are independent.

Under the assumption of independence we have that

$$P(Y = k, X = d) = P(X = d)P(Y = k) = p_k^Y p_d^X \tag{106}$$

If we have samples $\{s_i = (y_i, x_i)\}_{i=1}^N$ the expected number of observations of $Y = k$ and $X = d$ are

$$E_{k,d} = N p_k^Y p_d^X \tag{107}$$

If p^Y and p^X were known, we could use the test in section 4. Whoever, we assume we do not know them and must estimate them from the data using the methods in section **??**. We then loose one degree of freedom per category on X, and category on Y.

The χ^2 test of independence

$$C^2 = \sum_{k=1}^{K} \sum_{d=1}^{D} \frac{(\hat{O}_{k,d} - \hat{E}_{k,d})^2}{\hat{E}_{k,d}} \tag{108}$$

has a $\chi^2_{(K-1)\times(D-1)}$ distribution. As a χ^2_ν distribution becomes more and more concentrated as ν increases, this test is less powerfol that the one in section 4 were we would have had $K \times D - 1$ degrees of freedom.

In equation 108 we have

$$\hat{O}_{k,d} = \sum_{i=1}^{N} \mathbb{1}(y_i = d, x_i = k) \tag{109}$$

and

$$\hat{E}_{k,d} = N \hat{p}_k^Y \hat{p}_d^X \tag{110}$$

with $\hat{p}^Y$ and $\hat{p}^X$ been the MLE of X and Y probability distributions.

The test is particularly convenient to implement if we introduce the one-hot encodings for X and Y:

- $Z_d^X = 1$ if an only if $X = A_d$,
- $Z_k^Y = 1$ if an only if $Y = B_k$.

With those definitions, and given a set of observations, $z_{i,k}^Y$, $z_{i,d}^X$ for $i = 1, \dots N$

$$\hat{O}_{k,d} = \sum_i z_{i,k}^Y \, z_{i,d}^X = (Z^Y)^{\mathrm{T}} Z^X \tag{111}$$

where the last product is matrix multiplication and

$$\hat{p}_k^Y = \frac{1}{N}\sum_i z_{i,k}^Y = \frac{\hat{N}_k^Y}{N}$$

$$\hat{p}_d^X = \frac{1}{N}\sum_i z_{i,d}^X = \frac{\hat{N}_d^X}{N}. \tag{112}$$

$$\tag{113}$$

Using the notation on sufficient statistics $\hat{N}_k$ as defined in section 3.2.1, equation 69

5 Extreme Value Distribution

When performing Machine Learning procedures we will face many times the situation in which we will have have to pick up the best among a series of alternatives. This naturally introduces some bias, as even all the methods are equivalent, random fluctuations will make one of the alternatives look better.

To clarify this situation we analyze briefly the situation for Gaussian variables:

Lets assume we have x_c $\mathbb{N}(0,1)$ independent Normally distributed random variables with zero mean and unit standard deviation for $c = 1, \ldots, C$. We are interested on the distribution of

$$M_C = \max_c(x_1, x_2, \ldots, x_C) \tag{114}$$

Because the variables are independent

$$P(M_C < z) = \prod_c P(x_c < z) = \Phi(z)^C \tag{115}$$

where Φ is the cumulative density functions (CDF) of the normal distribution.

It can be show that

$$\min_{C\to\infty} P(M_C < z) = e^{e^{-\left(b_C + \frac{z}{b_C}\right)}} \tag{116}$$

with

$$b_C = \sqrt{2\log C} - \frac{\frac{1}{2}\log\log C + \log(2\sqrt{\pi})}{\sqrt{2\log C}} + o\left((logC)^{-\frac{1}{2}}\right) \tag{117}$$

The function

$$G(x) = e^{e^{-x}} \tag{118}$$

is known as the Gumbel or Extreme value distribution. It appears naturally as the limiting distribution of the maximum of independent random variables.

The implications of this asymptotical formula is that, the maximum of C Gaussian variables grows like the square root of the log of C. For $C \approx 1000$ $b_C \approx 3.11$

References

[1] Kevin P.Murphy. Conjugate bayesian analysis of the gaussian distribution. 2007. http://www.cs.ubc.ca/~murphyk/Papers/bayesGauss.pdf.

[2] Kevin P. Murphy. *Machine Learning, A Probabilistic Approach*. The MIT Press, 2012.

[3] David R. Hunter. *Statistics 553: Assumptotic Tools*. Self Published, 2006. http://sites.stat.psu.edu/~drh20/asymp/fall2006/lectures/ANGELchpt07.pdf.