MLE/MAP assignment

Suppose we have a probability distribution or density $p(x;\theta)$, where x may be discrete or continuous depending on the problem we are interested in. θ specifies the parameters of this distribution such as the mean and the variance of a one dimensional Gaussian. Different settings of the parameters imply different distributions over x. The available data, when interpreted as samples $x_1, \ldots, x_n$ from one such distribution, should favor one setting of the parameters over another. We need a formal criterion for gauging how well any potential distribution $p(\cdot|\theta)$ "explains" or "fits" the data. Since $p(x|\theta)$ is the probability of reproducing any observation x, it seems natural to try to maximize this probability. This gives rise to the Maximum Likelihood estimation criterion for the parameters θ:

$$\hat{\theta}_{\mathrm{ML}} = \arg\max_{\theta} L(x_1, \ldots, x_n; \theta) = \arg\max_{\theta} \prod_{i=1}^{n} p(x_i|\theta) \tag{1}$$

where we have assumed that each data point x_i is drawn independently from the same distribution so that the likelihood of the data is $L(x_1, \ldots, x_n; \theta) = \prod_{i=1}^{n} p(x_i; \theta)$. Likelihood is viewed primarily as a function of the parameters, a function that depends on the data. The above expression can be quite complicated (depending on the family of distributions we are considering), and make maximization technically challenging. However, any monotonically increasing function of the likelihood will have the same maxima. One such function is log-likelihood $\log L(x_1, \ldots, x_n; \theta)$; taking the log turns the product into a sum, making derivatives significantly simpler. We will maximize the log-likelihood instead of likelihood.

Problem 1: Maximum Likelihood Estimation

Consider a sample of n real numbers $x_1, x_2, \ldots, x_n$ drawn independently from the same distribution that needs to be estimated. Assuming that the underlying distribution belongs to one of the following parametrized families, the goal is to estimate its parameters (each family should be treated separately):

$$\text{Uniform}: \quad p(x; a) = \frac{1}{a} \text{ for } x \in [0, a],\ 0 \text{ otherwise} \tag{2}$$

$$\text{Exponential}: \quad p(x; \eta) = \frac{1}{\eta} \exp(-x/\eta),\ \eta > 0 \tag{3}$$

$$\text{Gaussian}: \quad p(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{4}$$

1. Derive the maximum likelihood estimators $\hat{a}_{\mathrm{ML}}$, $\hat{\eta}_{\mathrm{ML}}$, $\hat{\mu}_{\mathrm{ML}}$, $\hat{\sigma}^2_{\mathrm{ML}}$. The estimators should be obtained by maximizing the log-likelihood of the dataset under each of the families, and should be a function of $x_1, x_2, \ldots, x_n$ only.

To assess how well an estimator $\hat{\theta}$ recovers the underlying value of the parameter θ, we study its *bias* and *variance*. The bias is defined by the expectation of the deviation from the true value under the true distribution of the sample $(X_1, X_2, \ldots, X_n)$:

$$\mathrm{bias}(\hat{\theta}) = \mathrm{E}_{X_i \sim P(X|\theta)} \left[\hat{\theta}(X_1, X_2, \ldots, X_n) - \theta\right] \tag{5}$$

Biased (i.e. with a non-zero bias) estimators systematically under-estimate or over-estimate the parameter.

The variance of the estimator

$$\operatorname{var}(\hat{\theta}) = \operatorname{E}_{X_i \sim P(X|\theta)}\left[\left(\hat{\theta}(X_1, X_2, \ldots, X_n) - \operatorname{E}\left[\hat{\theta}(X_1, X_2, \ldots, X_n)\right]\right)^2\right] \tag{6}$$

measures the anticipated uncertainty in the estimated value due to the particular selection $(x_1, x_2, \ldots, x_n)$ of the sample. Note that the concepts of bias and variance of estimators are similar to the concepts of structural and approximation errors, respectively.

Estimators that minimize both bias and variance are preferred, but typically there is a trade-off between bias and variance.

2. Show that $\hat{a}_{\mathrm{ML}}$ is biased (no need to compute the actual value of the bias), $\hat{\eta}_{\mathrm{ML}}$ and $\hat{\mu}_{\mathrm{ML}}$ are unbiased.

3. Show that $\hat{\sigma}^2_{\mathrm{ML}}$, equal to the sample variance $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$, is biased. Show that the ML estimator of the variance becomes unbiased after multiplication with $n/(n-1)$. Let $\hat{\sigma}^2_{n-1}$ be this new estimator.

4. A standard way to balance the tradeoff between bias and variance is to choose estimators of lower *mean squared error*: $\operatorname{MSE}(\hat{\theta}) = \operatorname{E}_{X_i \sim P(X|\theta)}\left[(\hat{\theta} - \theta)^2\right]$. Show that $\operatorname{MSE}(\hat{\theta}) = \operatorname{bias}(\hat{\theta})^2 + \operatorname{var}(\hat{\theta})$ and that $\operatorname{MSE}(\hat{\sigma}^2_{\mathrm{ML}}) < \operatorname{MSE}(\hat{\sigma}^2_{n-1})$ even though $\hat{\sigma}^2_{\mathrm{ML}}$ is biased.

Problem 2: Maximum A-Posteriori Estimation

We want to determine the bias of an unfair coin for "heads" or "tails" from observing the outcome of a series of tosses. We model the coin by a single parameter θ that represents the probability of tossing heads.

Given n independent observed tosses $\mathcal{D} = \{x_1, \ldots, x_n\}$ out of which n_H are "heads", the likelihood function is:

$$p(\mathcal{D}|\theta) = \theta^{n_H}(1-\theta)^{n-n_H} \tag{7}$$

1. Show that $\hat{\theta}_{\mathrm{ML}} = n_H/n$. Thus if we toss the coin only once and we see "tails" ($n = 1$ and $n_H = 0$), according to maximum likelihood flipping the coin should always result in "tails".

While the maximum likelihood estimator is accurate on large training samples, if data is very scarce the estimated value is not that meaningful (on small samples the variance of the estimator is very high and it overfits easily). In contrast, in **Maximum A-Posteriori (MAP) estimation** we compensate for the lack of information due to limited observations with an *a priori* preference on the parameters based on prior knowledge we might have. In the case of the coin toss for instance, even without seeing any tosses we can assume the coin should be able to show both "heads" and "tails" ($\theta \neq 0$).

We express the prior preference/knowledge about θ by a distribution $p(\theta)$ (the *prior*). Assuming that θ and the observed sample are characterized by an underlying joint probability $p(\theta, \mathcal{D})$, we can use the Bayes rule to express our adjusted belief about the parameters after observing the trials (the *posterior*):

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} \tag{8}$$

where $p(\mathcal{D}) = \int p(\mathcal{D}|\theta')p(\theta')d\theta'$ normalizes the posterior. Maximization of the posterior distribution gives rise to the *Maximum A-Posteriori* (MAP) estimate of the parameters:

$$\hat{\theta}_{\mathrm{MAP}} = \arg\max_{\theta} p(\theta|\mathcal{D}) = \arg\max_{\theta} p(\mathcal{D}|\theta)p(\theta) \tag{9}$$

As in maximum likelihood, to compute the MAP estimate it is often easier to maximize the logarithm $\log p(\theta) + \log p(\mathcal{D}|\theta)$.

For the coin toss we will consider separately each of the following priors:

$$\text{Discrete}: \quad p^1(\theta) = \begin{cases} 0.5 & \text{if } \theta = 0.5 \\ 0.5 & \text{if } \theta = 0.4 \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

$$\text{Beta}: \quad p^2(\theta) = \frac{1}{Z}\theta^{\alpha-1}(1-\theta)^{\beta-1} \tag{11}$$

Here α and β are *hyperparameters* that should be given, not estimated, and Z is a normalization constant needed to make $p^2(\theta)$ integrate to 1 whose actual value is not important.

2. Prior $p^1(\theta)$ translates into a strong belief that the coin is either fair, or biased towards "tails" with a "heads" probability of 0.4. Express the MAP estimate $\hat{\theta}^1{}_{\text{MAP}}$ under this prior as a function of n_H/n.

3. The Beta prior expresses the belief that θ is likely to be near $\alpha/(\alpha+\beta)$. The larger $\alpha+\beta$ is, the more peaked the prior, and the stronger the bias that θ is close to $\alpha/(\alpha+\beta)$. Derive $\hat{\theta}^2{}_{\text{MAP}}$ under the Beta prior and show that when n approaches infinity the MAP estimate approaches the ML estimate, thus the prior becomes irrelevant given a large number of observations.

4. Compare qualitatively $\hat{\theta}^1{}_{\text{MAP}}$ and $\hat{\theta}_{\text{ML}}$. Assuming that the coin has a true "heads" probability of 0.41, which of the two estimators is likely to learn it faster? If data is sufficient, which of the two estimators is better?

MLE and MAP on Gaussian again

In this problem we will find the maximum likelihood estimator (MLE) and maximum a posteriori (MAP) estimator for the mean of a univariate normal distribution. Specifically, we assume we have N samples, $x_1, \ldots, x_N$ independently drawn from a normal distribution with *known* variance σ^2 and *unknown* mean μ.

1. Please derive the MLE estimator for the mean μ. Make sure to show all of your work.

2. Now derive the MAP estimator for the mean μ. Assume that the prior distribution for the mean is itself a normal distribution with mean ν and variance β^2. Please show all of your work. HINT: You may want to make use of the fact that:

$$\beta^2\left(\sum_{i=1}^{N}(x_i-\mu)^2\right)+\sigma^2(\mu-\nu)^2 = \left[\mu\sqrt{N\beta^2+\sigma^2}-\frac{\sigma^2\nu+\beta^2\sum_{i=1}^{N}x_i}{\sqrt{N\beta^2+\sigma^2}}\right]^2-\frac{[\sigma^2\nu+\beta^2\sum_{i=1}^{N}x_i]^2}{N\beta^2+\sigma^2}+\beta^2\left(\sum_{i=1}^{N}x_i^2\right)+\sigma^2\nu^2$$

3. Please comment on what happens to the MLE and MAP estimators as the number of samples N goes to infinity.

4.parameter estimation

The **Poisson distribution** is a useful discrete distribution which can be used to model the number of occurrences of something per unit time. For example, in networking, packet arrival density is often modeled with the Poisson distribution. That is, if we sit at a computer, count the number of packets arriving in each time interval, say every minute, for 30 minutes, and plot the histogram of how many time intervals had X number of packets, we expect to see something like a Poisson PMF curve.

If X (e.g. packet arrival density) is Poisson distributed, then it has PMF:

$$P(X|\lambda) := \frac{\lambda^X e^{-\lambda}}{X!},$$

where $\lambda > 0$ is the parameter of the distribution and $X \in \{0, 1, 2, \ldots\}$ is the discrete random variable modeling the number of events encountered per unit time.

Part A: Derive the expression for log likelihood

It can be shown that the parameter λ is the **mean** of the Poisson distribution. In this part, we will estimate this parameter from the number of packets observed per unit time $X_1, ..., X_n$ which we assume are drawn i.i.d from $Poisson(\lambda)$.

1. Recall that the *bias* of an estimator of a parameter θ is defined to be the difference between the expected value of the estimator and θ.

   (a) Show that $\hat{\lambda} = \frac{1}{n}\sum_i X_i$ is the maximum likelihood estimate of λ.

(b) Show that it is unbiased (that is, show that $E[\hat{\lambda}] - \lambda = 0$). Recall that $E[a + b] = E[a] + E[b]$ (linearity of expectations).

2. Now let's be Bayesian and put a prior distribution over the parameter λ.

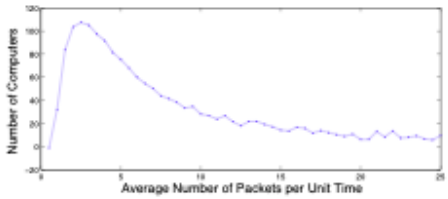

Figure 1: Just giving you some motivation. Don't take it so seriously.

Your friend in networking hands you a typical plot showing the counts of computers at a university cluster with different average packet arrival densities (Figure 1). Your extensive experience in statistics tells you that the plot resembles a Gamma distribution pdf. So you believe a good prior distribution for λ may be a Gamma distribution. Recall that the Gamma distribution has pdf:

$$P(\lambda|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}, \quad \lambda > 0$$

Also, if $\lambda \sim \Gamma(\alpha, \beta)$, then it has mean α/β and the mode is $(\alpha - 1)/\beta$ for $\alpha > 1$.[2]

Assuming that λ is distributed according to $\Gamma(\lambda|\alpha, \beta)$, compute the posterior distribution over λ.

Hint:

$$\lambda^{\sum X_i + \alpha - 1} e^{-\lambda(n+\beta)}$$

looks like a Gamma distribution! Is the rest of the expression constant with respect to λ?

3. Derive an analytic expression for the maximum a posteriori (MAP) estimate of λ under a $\Gamma(\alpha, \beta)$ prior.

4. Given N samples $x_1, x_2, \ldots, x_N$ drawn independently from a Gaussian distribution with variance σ^2 and unknown mean μ, find the MLE of the mean.

   (a) $\mu_{MLE} = \frac{\sum_{i=1}^{N} x_i}{\sigma^2}$

   (b) $\mu_{MLE} = \frac{\sum_{i=1}^{N} x_i}{2\sigma^2 N}$

   (c) $\mu_{MLE} = \frac{\sum_{i=1}^{N} x_i}{N}$

   (d) $\mu_{MLE} = \frac{\sum_{i=1}^{N} x_i}{N-1}$

5. Continuing with the above question, assume that the prior distribution of the mean is also a Gaussian distribution, but with parameters mean μ_p and variance σ_p^2. Find the MAP estimate of the mean.

   (a) $\mu_{MAP} = \frac{\sigma^2 \mu_p + \sigma_p^2 \sum_{i=1}^{N} x_i}{\sigma^2 + N\sigma_p^2}$

   (b) $\mu_{MAP} = \frac{\sigma^2 + \sigma_p^2 \sum_{i=1}^{N} x_i}{\sigma^2 + \sigma_p^2}$

   (c) $\mu_{MAP} = \frac{\sigma^2 + \sigma_p^2 \sum_{i=1}^{N} x_i}{\sigma^2 + N\sigma_p^2}$

   (d) $\mu_{MAP} = \frac{\sigma^2 \mu_p + \sigma_p^2 \sum_{i=1}^{N} x_i}{N(\sigma^2 + \sigma_p^2)}$

6. Which among the following statements is (are) true?

   (a) MAP estimates suffer more from overfitting than maximum likelihood estimates.

   (b) MAP estimates are equivalent to the ML estimates when the prior used in the MAP is a uniform prior over the parameter space.

   (c) One drawback of maximum likelihood estimation is that in some scenarios (hint: multinomial distribution), it may return probability estimates of zero.

   (d) The parameters which minimize the expected Bayesian L1 Loss is the median of the posterior distribution.