Tina Memo No. 2011-003
Internal Report

Quantitative Counting with Bayes Theorem

P. Tar [1]

Last updated
9 / 6 / 2011

ISBE, Medical School,
University of Manchester,
Stopford Building, Oxford Road,
Manchester, M13 9PT, UK.

[1] Supervisor N.A. Thacker.

Quantitative Counting with Bayes Theorem

Abstract

Given a dataset generated by various processes, an estimate of the quantity of data points associated with each generation process can be achieved using a probabilistic count of evidence using Bayes Theorem. However, the usefulness of these counts is limited by the presence of uncertainties stemming from the determination of quantities using maximum likelihood and the imprecision of density estimates. Systematic effects are also present in practice due to a lack of agreement between modelled data densities and actual data. This document follows on from Tina memo 2010-008 and shows how predictable sources of variation can be computed and how systematic effects can be minimised by using mixtures of density estimates to improve model-data agreement.

1 Quantitative counting

Standard classifiers (Support Vector Machines, Random Forests, Boosting etc.) can be used to give decisive class labels to incoming data points. A discrete count of resulting labels could form the basis of an estimate of the quantities of classes present. However, any ambiguity will inevitably result in classification mistakes and potentially bias results. An alternative method can weight the counting of data points by the probability that they belong to a given class, rather than forcing a hard decision. Assuming honest probabilities this method should correct for misclassifications thereby providing an unbiased estimate of quantities. Given a set of data generation processes, $J = \{1, 2, \ldots, N\}$, the quantity of any given class occurring in a region can be estimated using

$$Q(j) = \sum_{d \in R} P(j|X_d)$$

where X_d is evidence observed at location d within region R of the dataset and $P(j|X_d)$ is the probability class j was the source of the observation. The class probability can be found using Bayes Theorem

$$P(j|X) = \frac{P(X|j)Q(j)}{\sum_{l \in J} P(X|l)Q(l)}$$

where $P(X|j)$ is a density estimate for class j and $Q(j)$ is the amount of class j instances present in the region. Maximum likelihood estimates for the values of $Q(j)$ are those which maximise the probability modelled data densities account for the actual observations. These can be found by minimising

$$F = \sum_{d \in R} -\log \sum_{l \in J} P(X_d|l)Q(l)$$

which can be achieved using the parameter update methods for Expectation Maximisation.

The usefulness of the resulting quantities will be limited by their accuracy. To be used quantitatively an estimate of their uncertainty is needed. The remainder of this document addresses the issue of uncertainty and how it can be managed.

2 Calculating covariances

Summary outputs presented to end users can only be meaningfully interpreted if they are accompanied by estimates of their errors, e.g. points on plots should be overlaid with error bars denoting clear confidence intervals. This is especially important if results are to be used in further calculations. A covariance matrix provides all necessary information for both interpretation and further processing therefore determining the covariance matrix of quantitative counts, $Q(j)$, is essential.

A covariance matrix for quantitative counting must take into consideration the most significant sources of uncertainty. These sources include independent contributions from the variance in maximum likelihood determination of quantities and the imprecision of density estimates. Elements of the covariance matrix are then given by

$$\mathbf{C}_{ij} = cov_{CRB}(Q(i), Q(j)) + cov_{DE}(Q(i), Q(j))$$

where cov_{CRB} is the contribution from the likelihood, which can be estimated using the Cramer Rao Bound, and cov_{DE} is the contribution from imprecise density estimates, which can be estimated using error propagation. Contributions from these two sources are derived separately in the following two sections.

2.1 Uncertainty in counting

The contribution to the uncertainty of counts due to their maximum likelihood determination can be estimated using the Cramer Rao lower variance bound

$$\frac{\partial^2 F}{\partial Q(i) \partial Q(j)} \geq \frac{1}{cov(Q(i), Q(j))}$$

Computing the first and second order derivatives yields

$$\frac{\partial F}{\partial Q(j)} = -\sum_{d \in R} \frac{P(X_d|j)}{\sum_{l \in J} P(X_d|l) Q(l)}$$

$$\frac{\partial^2 F}{\partial Q(i) \partial Q(j)} = \sum_{d \in R} \frac{P(X_d|i) P(X_d|j)}{[\sum_{l \in J} P(X_d|l) Q(l)]^2}$$

The similarity with Bayes Theorem allows the second derivative to be formulated as

$$\frac{\partial^2 F}{\partial Q(i) \partial Q(j)} = \frac{1}{Q(i) Q(j)} \sum_{d \in R} P(i|X_d) P(j|X_d)$$

So that the final inverse covariance estimate is

$$C_{ij}^{-1} \approx \frac{\sum_{d \in R} P(i|X_d) P(j|X_d)}{Q(i) Q(j)}$$

from which the covariance matrix can be computed using a standard matrix inversion algorithm. The covariance contribution from this lower bound shall be referred to as $cov_{CRB}(Q(i), Q(j))$. The next section will derive covariance contributions due to statistical perturbations in density estimates.

2.2 Uncertainty in density estimation

Uncertainty due to imprecision in density estimation can be tracked using error propagation. The characteristics of these errors will be dependent upon the type of density models used. The least sophisticated method of density estimation is that based upon histograms. Each bin, i.e. each pattern X, within a histogram based density will be subject to independent Poisson errors. The following contribution to the covariance matrix assumes the use of such histograms.

Patterns are likely to occur multiple times within a dataset with each occurrence of a single pattern being subject to the same error. Because of this correlation it is better to formulate quantities in terms of unique patterns rather than data points giving

$$Q(j) = \sum_{X \in R} \frac{P(X|j) Q(j)}{P(X|j) Q(j) + P(X|\bar{j}) Q(\bar{j})} Q(X)$$

where $Q(X)$ is the total quantity of pattern X appearing within region R and the denominator in Bayes Theorem has been separated into independent error components where $\bar{j}$ represents all classifications other than j. The covariance is then given by

$$cov_{DE}(Q(i),Q(j)) = \sum_{X\in R}\left[\sum_{k\in J}\sum_{l\in J}\left(\frac{\partial Q(i)}{\partial P(X|k)}\right)\left(\frac{\partial Q(j)}{\partial P(X|l)}\right)cov(P(X|k),P(X|l))\right]Q(X)^2$$

where $cov_{DE}(P(X|k),P(X|l))$ is equal to zero when $k \neq l$, due to independence, allowing the covariance to be expressed as

$$cov_{DE}(Q(i),Q(j)) = \sum_{X\in R}\left[\sum_{k\in J}\left(\frac{\partial Q(i)}{\partial P(X|k)}\right)\left(\frac{\partial Q(j)}{\partial P(X|k)}\right)var(P(X|k))\right]Q(X)^2$$

The derivative of a quantity with respect to a density estimate is dependent on the relationship between the class for which a quantity is sought and the class of the density estimate. If the quantity is for the same class as the density, then that density estimate will appear in both the numerator and denominator in Bayes Theorem. If the quantity is of a different class than the density, then the density estimate will only appear in the denominator. The derivative with respect to a computed quantities' own class density estimate is given by

$$\Delta_{jj} = \frac{\partial Q(j)}{\partial P(X|j)} = \frac{Q(j)[P(X|j)Q(j) + P(X|\bar{j})Q(\bar{j})] - P(X|j)Q(j)^2}{[P(X|j)Q(j) + P(X|\bar{j})Q(\bar{j})]^2}$$

$$= \frac{P(X|\bar{j})Q(\bar{j})Q(j)}{[P(X|j)Q(j) + P(X|\bar{j})Q(\bar{j})]^2}$$

$$= \frac{P(X|\bar{j})Q(\bar{j})Q(j)}{[\sum_{l\in J} P(X|l)Q(l)]^2}$$

If the modelled data density matches the actual data then $\sum_{l\in J} P(X|l)Q(l) = Q(X)$. From this and Bayes Theorem the derivative can be simplified to

$$\Delta_{jj} = \frac{P(\bar{j}|X)Q(j)}{Q(X)}$$

The derivative with respect to other class density estimates is given by

$$\Delta_{jk} = \frac{\partial Q(j)}{\partial P(X|k)} = \frac{-P(X|j)Q(j)Q(k)}{[P(X|j)Q(j) + P(X|\bar{j})Q(\bar{j})]^2}$$

$$= \frac{-P(X|j)Q(j)Q(k)}{[\sum_{l\in J} P(X|l)Q(l)]^2}$$

which simplifies to

$$\Delta_{jk} = \frac{-P(j|X)Q(k)}{Q(X)}$$

The short-hand Δ_{jk}, meaning the partial derivative of class quantity j with respect to class density estimate k, will be used subsequently for compactness. There are four derivative permutations when computing the covariance giving

$$cov_{DE}(Q(i),Q(j)) = \sum_{X\in R}\left[\sum_{k=i=j}\Delta_{kk}\Delta_{kk}var(P(X|k))Q(X)^2 + \sum_{k\neq i,k\neq j}\Delta_{ik}\Delta_{jk}var(P(X|k))Q(X)^2\right.$$
$$\left. + \sum_{k=i,k\neq j}\Delta_{kk}\Delta_{jk}var(P(X|k))Q(X)^2 + \sum_{k\neq i,k=j}\Delta_{ik}\Delta_{kk}var(P(X|k))Q(X)^2\right]$$

Assuming the histograms from which density estimates were computed were constructed using a total of $T(k)Q(k)$ training examples (i.e. the training used $T(k)$ times as much data as is being analysed) the variance of the estimates is given by

$$var(P(X|k)) = \frac{P(X|k)}{T(k)Q(k)}$$

The four components of the covariance calculation are now presented starting with

$$\sum_{k=i=j} \Delta_{kk}\Delta_{kk} var(P(X|k))Q(X)^2$$

$$= \sum_{k=i=j} \frac{P(\bar{k}|X)Q(k)}{Q(X)} \times \frac{P(\bar{k}|X)Q(k)}{Q(X)} \times \frac{P(X|k)}{T(k)Q(k)} \times Q(X)^2$$

$$= \sum_{k=i=j} \frac{P(\bar{k}|X)^2 P(X|k)Q(k)}{T(k)}$$

And the next term

$$\sum_{k\neq i, k\neq j} \Delta_{ik}\Delta_{jk} var(P(X|k))Q(X)^2$$

$$= \sum_{k\neq i, k\neq j} \frac{-P(i|X)Q(k)}{Q(X)} \times \frac{-P(j|X)Q(k)}{Q(X)} \times \frac{P(X|k)}{T(k)Q(k)} \times Q(X)^2$$

$$= \sum_{k\neq i, k\neq j} \frac{P(i|X)P(j|X)P(X|k)Q(k)}{T(k)}$$

And the next

$$\sum_{k=i, k\neq j} \Delta_{kk}\Delta_{jk} var(P(X|k))Q(X)^2$$

$$= \sum_{k=i, k\neq j} \frac{P(\bar{k}|X)Q(k)}{Q(X)} \times \frac{-P(j|X)Q(k)}{Q(X)} \times \frac{P(X|k)}{T(k)Q(k)} \times Q(X)^2$$

$$= \sum_{k=i, k\neq j} \frac{-P(\bar{k}|X)P(j|X)P(X|k)Q(k)}{T(k)}$$

And finally

$$\sum_{k\neq i, k=j} \Delta_{ik}\Delta_{kk} var(P(X|k))Q(X)^2$$

$$= \sum_{k\neq i, k=j} \frac{-P(i|X)Q(k)}{Q(X)} \times \frac{P(\bar{k}|X)Q(k)}{Q(X)} \times \frac{P(X|k)}{T(k)Q(k)} \times Q(X)^2$$

$$= \sum_{k\neq i, k=j} \frac{-P(i|X)P(\bar{k}|X)P(X|k)Q(k)}{T(k)}$$

When substituted back into the covariance calculation these terms are summed over patterns, X. For implementation purposes this would be more convenient if summed over data points, X_d. This can be achieved by dividing by $Q(X)$. As each covariance term contains a $P(X|k)Q(k)$ this division presents another opportunity to reformulate in terms of Bayes Theorem as $P(X|k)Q(k)/Q(X) = P(k|X)$ giving

$$cov_{DE}(Q(i),Q(j)) = \sum_{d\in R}\left[\sum_{k=i=j}\frac{P(\bar{k}|X_d)^2P(k|X_d)}{T(k)} + \sum_{k\neq i,k\neq j}\frac{P(i|X_d)P(j|X_d)P(k|X_d)}{T(k)}\right.$$
$$\left. + \sum_{k=i,k\neq j}\frac{-P(\bar{k}|X_d)P(j|X_d)P(k|X_d)}{T(k)} + \sum_{k\neq i,k=j}\frac{-P(i|X_d)P(\bar{k}|X_d)P(k|X_d)}{T(k)}\right]$$

Bringing the sum over all data points into each component and using $\sum_{d\in R}P(k|X_d) = Q(k)$ this can be more easily implemented by reusing quantities precomputed during counting giving

$$cov_{DE}(Q(i),Q(j)) = \left[\sum_{k=i=j}\frac{Q(k)}{T(k)}\sum_{d\in R}P(\bar{k}|X_d)^2\right] + \left[\sum_{k\neq i,k\neq j}\frac{Q(k)}{T(k)}\sum_{d\in R}P(i|X_d)P(j|X_d)\right]$$
$$+ \left[\sum_{k=i,k\neq j}\frac{Q(k)}{T(k)}\sum_{d\in R}-P(\bar{k}|X_d)P(j|X_d)\right] + \left[\sum_{k\neq i,k=j}\frac{Q(k)}{T(k)}\sum_{d\in R}-P(i|X_d)P(\bar{k}|X_d)\right]$$

An implementation which records values during training and counting will have all covariance terms readily available.

Together cov_{CRB} plus cov_{DE} should account for the majority of the statistical errors on computed quantities. However, this estimate assumes probability densities appropriately match data under analysis. If they do not match then systematic effects will be introduced, which are addressed next.

3 Systematic errors

The covariance matrix derived in the previous section assumes that modelled data densities, $\sum_{l\in J}P(X|l)Q(l)$, match actual data, $Q(X)$, within statistical margins of error. However, in practice, a non-trivial dataset containing complex stochastic patterns is unlikely to reflex models perfectly leading to systematic effects caused by a lack of model-data agreement. This mismatch will impact upon any probabilities computed using Bayes Theorm

$$P(j|X) = \frac{P(X|j)Q(j)}{\sum_{l\in J}P(X|l)Q(l)}$$

The effect upon $P(j|X)$ due to changes (i.e. errors) in density estimate $P(X|j)$ is given by

$$\frac{\partial P(j|X)}{\partial P(X|j)} = \frac{Q(j)\sum_{k\neq j}P(X|k)Q(k)}{Q(X)^2} = \frac{Q(j)P(\bar{j}|X)}{Q(X)}$$

The systematic error introduced by a mismatch between model and data which is beyond statistical fluctuations can be estimated by

$$P(j|X) - P'(j|X) = \left(\frac{\partial P(j|X)}{\partial P(X|j)}\right)(P(X|j) - P'(X|j))$$

Taking the expectation of this difference provides a means of assessing the impact of systematic errors upon computed quantities, $Q(j)$, giving a quality function

$$S(Q(j)) = E < P(j|X) - P'(j|X) >= \sum_X Q(X)P(j|X)\left(\frac{\partial P(j|X)}{\partial P(X|j)}\right)(P(X|j) - P'(X|j))$$

which is summed over patterns, X, which have correlated errors as described previously. This can be converted to a sum over data points giving

$$S(Q(j)) = Q(j)(\bar{j})\sum_d P(j|X_d)P(\bar{j}|X_d)\frac{P(X_d|j) - P'(X_d|j)}{Q(X_d)}$$

For fixed $P(X_d|j)$, $P'(X_d|j)$ and $Q(\bar{j})/Q(X_d)$ ratio, we can see that $S(Q(j)) \propto Q(j)$ giving an error which grows proportionally as quantities, $Q(j)$, increase. This is in contrast to the statistical errors computed earlier which scale as $\sqrt{Q(j)}$. As P' is unknown there is no way to compute the actual systematic error without resorting to Monte-Carlos, which may be impactical. To trust any confidence intervals the quantity of data analysed must be small enough so statistical errors dominates, whilst at the same time be sufficiently large for an analysis to serve a useful purpose. This trade-off can be improved by improving density estimates. One method which may improve this situation is presented next.

4 Mixture models

The density estimates, $P(X|j)$, were assumed to be simple histograms, one for each category of feature expected to be encountered. If these simple models do not match actual data then systematic effects will be introduced, as explained in the previous section. An improved model with greater flexibility could make use of a mixture of subclasses weighted to better match incoming data

$$P(X|K) = \sum_{k \in K} P(X|k)\alpha_k$$

where K is a set of subclasses describing the constituent parts of a wider category in which each component $k \in K$ is modelled separately then mixed. Each component can be scaled by α_k allowing the freedom to change the shape of density estimate $P(X|K)$ to improve model-data agreement. Incorporating this into the quantitative count gives

$$Q(K) = \sum_{d \in R} \frac{[\sum_{k \in K} P(X_d|k)\alpha_k]Q(K)}{\sum_{L \in J}[\sum_{l \in L} P(X_d|l)\alpha_l]Q(L)}$$

which illustrates the intention of the weighted mixture model but does not address the issue of how an algorithm can determine appropriate scaling factors. If an application is only concerned with final quantities then any scaling can be factored into sub-quantities giving

$$Q(K) = \sum_{k \in K} \sum_{d \in R} \frac{P(X_d|k)Q(k)}{\sum_{l \in J} P(X_d|l)Q(l)}$$

It can be seen that the covariance matrix for sub-quantities is exactly the same as was derived earlier. Now error propagation can be used to provide the variance on any mixture count

$$[var(Q(K))] = \mathbf{DCD}^T$$

where $\mathbf{C}$ is the covariance matrix of sub-quantities and $\mathbf{D}$ is a vector defining which subcomponents appear within K

$$\mathbf{D} = [\delta_{1 \in K}, \delta_{2 \in K}, \ldots, \delta_{N \in K}]$$

where $\delta_{i \in K}$ is 1 if subcomponent i belongs within the set K, 0 otherwise.

5 Summary

Given a set of classifications, $j \in J$, with distributions of patterns, $P(X|j)$, it is possible to count the quantities, $Q(j)$, of data points within a dataset belonging to each classification using a sum of probabilities computed via Bayes Theorem. Such a probabilistic count can be performed within an Expectation Maximisation algorithm which, upon convergence, provides the maximum likelihood solutions. To allow estimated quantities to be meaningfully interpreted the Cramer Rao bound can be applied to the likelihood function to give a lower limit for the associated covariance matrix. The distribution of patterns for each classification can be sampled using simple histograms yielding density estimates subject to Poisson errors. These errors will increase the covariances in a way which can be derived using error propagation. Together, the CRB and Poisson errors can be used to create confidence intervals within which quantities can be trusted. Unfortunately this will only generalise to other datasets properly if there is suitable agreement between density estimates and analysed data. Any mismatch will lead to systematic

errors which scale proportionally to the quantities of classes present, whereas the statistical errors increase more slowly like the square-root of the quantities. There will be a point at which quantities become large enough that systematic effects dominant the uncertainty in estimated values. This critical point can be increased with better density estimations. A mixture of sub-classes of density estimates can be linearly combined to create more flexible densities which can change shape to better fit incoming data. Error propagation can again be used to combine the covariances from subclass quantities into the variance of superordinate class quantities. Such mixture models may reduce systematic effects at the cost of increasing statistical variations, but this may be a necessary penalty for increasing the amount of data which can be meaningfully analysed within predictable errors.