MLE Notes

Dan Beatty

February 27, 2007

Introduction

- Maximum-Likelihood Estimation
- Example of a Specific Case
- The Gaussian Case unknown μ and σ Bias
- Appendix ML Problem Statement

[1]

- Data availability in a Bayesian framework
  - Optimal classifiers require knowledge on priors $P(\omega_i)$ and class-conditional densities $P(x|\omega_i)$:
  - Neither of which are completely available at the time of classification.
- Design a classifier from a training sample
  - No problem with prior estimation
  - These prior probabilities are used to calculated estimated class conditional densities $P(x|\omega_i)$. This follows from the definition of mathematical expectation which requires an infinite set of samples. Training samples are supposed to be small enough to not over condition the classifier, but the expected values are only rough mean, and rely on representation assumptions which may not be true.
  - Samples are often too small for class-conditional estimation (large dimension of feature space!)
- A priori information about the problem
- Normality of $P(\vec{x}|\omega_i)$, $P(\vec{x}|\omega_i)$ $N(\mu_i, \Sigma_i)$
  Characterized by 2 parameters

- Estimation techniques:

  - Maximum Likelihood (ML) and the Bayesian estimations
  - Results are nearly identical, but the approaches are different.

  [?]

- “Parameters in ML estimation are fixed but unknown!”

- ‘Best estimate parameters’ are obtained by maximizing the probability obtaining the samples observed.

- Bayesian methods view parameters as random variables as random variables having some known distribution. The technique leads to *Bayesian Learning*.

- In either approach, we use $P(\omega_i|\vec{x})$ for our classification rule!

1 Maximum-Likelihood Estimation

- A good maximum likelihood estimator has good convergence properties as the sample size increases.

- Simpler than any many other alternative techniques

- Assume classes c

- Data sets $\mathcal{D}_1, ..., \mathcal{D}_n$ is drawn from independent and identically distributed random variables obeying $P(x|\omega_j)$.

$$P(x|\omega_j)\ N(\mu_j, \Sigma_j) \tag{1}$$

$$P(x|\omega_j) \equiv P(x|\omega_j, \theta_j) \tag{2}$$

$$\vec{\theta}\ N(\mu_j, \Sigma_j) = (\mu_j^1, \mu_j^2 ... \sigma^{11}, \sigma^{22}, cov(\vec{x_j}^m, \vec{x_j}^n)) \tag{3}$$

- Use the information provided by the training samples to estimate θ s.t.

$$\theta = (\theta_1, \theta_2, ..., \theta_c)$$

  and each θ_i for $i = 1, 2, ..., c$ is associated with each category

- Suppose that $\mathcal{D}$ contains n samples, $x_1, x_2, ..., x_n$

$$P(\mathbf{D}|\vec{\theta}) = \prod_{k=1}^{k=n} P(\vec{x}_k|\vec{\theta})$$

  $P(\mathbf{D}|\vec{\theta})$ is called the likelihood of $\vec{\theta}$ w.r.t the set of samples.

- ML estimate of θ is by definition the value that maximizes $P(\mathbf{D}|\vec{\theta})$, and that value is denoted $\hat{\vec{\theta}}$.
- Optimal estimation for a number of parameters of p
  - Let $\vec{\theta} = (\theta_1, \theta_2, ..., \theta_p)^T$ and let ∇_θ be the gradient operator

$$\nabla_{\vec{\theta}} \equiv \begin{bmatrix} \frac{\partial}{\partial\theta_1} \\ \frac{\partial}{\partial\theta_2} \\ . \\ . \\ . \\ \frac{\partial}{\partial\theta_p} \end{bmatrix} \tag{4}$$

  - We define $l(\theta)$ as the log-likelihood function

$$l(\theta) = \ln P(D|\theta)$$

  - New problem statement: determine θ that maximizes the log-likelihood

$$\hat{\theta} = \arg\max_\theta l(\theta)$$

    Set of necessary conditions for an optimum is defined on the likelihood from data set $\mathcal{D}$. The equation is specified in equation 5, and the maximization of the estimate gradient occurs as stated in equation 6.

$$\nabla_\theta I = \sum_{k=1}^{k=n} \nabla_\theta \ln P(\vec{x_k}|\theta) \tag{5}$$

$$\nabla_\theta I \Rightarrow 0 \tag{6}$$

    So far this has been a derivation of probability with vector calculus. As with all converging results, the result is still simply an estimate on the limit. A related class of estimators is called maximum a posteriori (MAP). MAP estimators have a drawback in when prior probabilities change due to a non-linear transformation of the parameter space.

In both the case of unknown mean and unknown covariance, the mathematical expectation function comes into play to define the sample mean and sample covariance. This is not surprising since that is how these values are defined in the first place.

1.1 Specific Case Example: Unknown mean $\vec{\mu}$

- $P(x_i|\mu)\ N(\mu,\Sigma)$ (Samples are drawn from a multivariate normal population)

$$\ln p(\vec{x}_k|\vec{\mu}) = -\frac{1}{2}(\ln[(2\pi)^d|\Sigma|] + (\vec{x}_k - \vec{\mu})^T\Sigma^{-1}(\vec{x}_k - \vec{\mu})) \tag{7}$$

$$\nabla_u \ln p(\vec{(x)}|\vec{\mu}) = \Sigma^{-1}(\vec{x}_k - \vec{\mu}) \tag{8}$$

- $\theta = \mu$ therefore: the ML estimate for μ must satisfy:
- Derivations via multiplication by σ and rearranging, we obtain

$$\sum_{k=1}^{n} \Sigma^{-1}(\vec{x}_k - \hat{\vec{\mu}}) = 0 \tag{9}$$

$$\hat{\vec{\mu}} = \frac{1}{n}\sum_{k=1}^{n} x_k \tag{10}$$

  Just the arithmetic average of the samples of the training samples! This is not surprising as the definition of expected values is based on approximation of the mean, aka the sample mean.

- Conclusion: If $P(x_k|\omega_j)$ such that $j = 1, 2, ..., c$ is supposed to be Gaussian in d-dimensional feature space; then we can estimate the vector $\theta = (\theta_1, \theta_2, ..., \theta_c)^T$ and perform optimal classification!

1.2 Specific Case Example: Unknown mean and covariance (μ, Σ)

Univariate case for unknown μ, Σ yields

$$\ln p(x_k|\vec{\theta}) = -\frac{1}{2}\ln 2\pi\theta_2 - \frac{1}{2\theta_2}(x_k - \omega_1)^2 \tag{11}$$

$$\nabla_\theta l = \nabla_\theta \ln p(x_k|\vec{\theta}) = \begin{bmatrix} \frac{1}{\theta_2}(x_k - \theta_1) \\ -\frac{1}{2\theta_2} + \frac{(x_k-\theta_1)^2}{2\theta_2^2} \end{bmatrix} \tag{12}$$

$$\sum_{k=1}^{n} \frac{1}{\hat{\theta}_2}(x_k - \hat{\theta}_1) = 0 \tag{13}$$

$$-\sum_{k=1}^{n} \frac{1}{\hat{\theta}_2} + \sum_{k=1}^{n} \frac{(x_k - \hat{\theta}_1)^2}{\hat{\theta}_2^2} = 0 \tag{14}$$

To satisfy these conditions, θ_1, θ_2 must yield the following for $\hat{\mu}$ and $\hat{\sigma^2}$ which are the sample mean and variance.

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} x_k \tag{15}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{k=1}^{n}(x_k - \hat{\mu})^2 \tag{16}$$

Likewise in the multi-variate case. Sample variance is always slightly different than the true variance for the distribution from which a sample is from.

The MLE of the variance σ^2 is biased

$$E[\frac{1}{n}\sum_{n=1}^{n}(x_i - \bar{x})^2] = \frac{n-1}{n}\sigma^2 \neq \sigma^2 \tag{17}$$

Considering the univariate case, let μ and σ^2 be the mean and variance of the Gaussian

$$\sigma_n^2 = E[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2] \tag{18}$$

$$= \frac{1}{n-1}E[\sum_{i=1}^{n}\{(x_i - \mu) - (\hat{\mu} - \mu)\}^2] \tag{19}$$

$$\frac{1}{n-1}E[\sum_{i=1}^{n}\{(x_i - \mu) - 2(x_i - \mu)(\hat{\mu} - \mu) + (\bar{\mu} - \mu)^2\}] \tag{20}$$

$$\frac{1}{n-1}[\sum_{i=1}^{n}\{E(x_i - \mu) - 2E(x_i - \mu)(\hat{\mu} - \mu) + E(\bar{\mu} - \mu)^2\}] \tag{21}$$

$$\frac{1}{n-1}[\sum_{i=1}^{n}\{E(x_i - \mu) - 2E(x_i - \mu)(\hat{\mu} - \mu) + E(\bar{\mu} - \mu)^2\}] \tag{22}$$

$$\tag{23}$$

$$E[(x_i - \mu)(\hat{\mu} - \mu)] \tag{24}$$

$$= E[(x_i - \mu)(\frac{1}{n}\sum_{j=1}^{n} x_k - \mu)] \tag{25}$$

$$= E[(x_i - \mu)(\frac{x_i - \mu}{n} + \frac{1}{n}\sum_{k=1,k\neq i} n x_k - \mu)] \tag{26}$$

$$= E[\frac{1}{n}(x_i - \mu)^2] + E[\frac{1}{n}(x_i - \mu)(\sum_{k=1,k\neq i} x_k - (n-1)\mu)] \tag{27}$$

$$= \frac{1}{n}\sigma^2 + 0 = \frac{\sigma^2}{n} \tag{28}$$

$$\tag{29}$$

$$E[(x_i - \mu)(\hat{\mu} - \mu)] \tag{30}$$

$$= E[(x_i - \mu)(\frac{1}{n}\sum_{j=1}^{n}(x_j - \mu))] \tag{31}$$

$$E[(x_i-\mu)(\frac{x_i-\mu}{n}+\frac{1}{n}\sum_{k=1,k\neq i}^{n}x_k-\mu)] \tag{32}$$

$$E[\frac{(x_i-\mu)^2}{n}]+E[\frac{x_i-\mu}{n}\sum_{k=1,k\neq i}^{n}(x_k-(n-1)\mu)] \tag{33}$$

$$=\frac{\sigma^2}{n}+0=\frac{\sigma^2}{n}E[\hat{\mu}-\mu]=\frac{\sigma^2}{n}\sigma_n^2=\frac{1}{n+1}[\sigma^2-\frac{2}{n}\sigma^2+\frac{\sigma^2}{n}] \tag{34}$$

$$\frac{n-1}{n-1}\sigma^2=\sigma^2 \rightarrow \text{ unbiased} \tag{35}$$

Similarly

$$E(\hat{\mu}-\mu)=\frac{\sigma^2}{n} \tag{36}$$

$$\sigma_n^2=\frac{1}{n-1}[\sigma^2-\frac{2}{n}\sigma^2+\frac{\sigma^2}{n}] \tag{37}$$

$$\text{unbiased } \frac{n-1}{n-1}\sigma^2=\sigma^2 \tag{38}$$

$$P(x_1,...,x_n|\theta)=\prod_{k=1}^{n}\prod_{i=1}^{d}\theta_2^{2k_i}(1-\theta_i)^{1-x_{k_i}} \tag{39}$$

$$l(0)==\sum_{k=1}^{n}\sum_{i=1}^{d}x_{ki}\ln\theta_i+(1-x_{ki})\ln(1-\theta_i) \tag{40}$$

We know that $p(x|\omega_1)\ N(\mu,1)$ and we assume that $p(x|\omega_2)\ N(\mu,1)$. Imagine, however, that the true underlying distribution is $p(x|\omega_2)\ N(1,10^5)$.

The MLE derivation is different. The unbiased estimator for Σ is given by equation 3-21 (page 90). Where C is the sample covariance matrix and $\hat{\Sigma}$ which equals:

$$\hat{\Sigma}=\frac{n-1}{n}C$$

Consider the problem of learning the mean of a univariate normal distribution from equations 34 and 35 we have

$$\mu_n=\frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2}m_n+\frac{\sigma^2}{n\sigma_0^2+\sigma^2}\mu_0 \tag{41}$$

$$\sigma_n^2=\frac{\sigma_0^2\sigma^2}{n\sigma_0^2+\sigma^2} \tag{42}$$

where $m_n=t_n$

μ_0 is formed by averaging n_0 fictitious samples x_k from $k = -n_0 + 1, -n_0 + 2, ..., 0$

$$\mu_0 = \frac{1}{n_0} \sum_{k=-n_0+1}^{0} x_k \tag{43}$$

$$\mu_n = \frac{\sum_{k=1}^{n} x_k}{n + \frac{\sigma^2}{\sigma_0^2}} + \frac{\frac{\sigma^2}{\sigma_0^2}}{\frac{\sigma^2}{\sigma_0^2} + n} \frac{1}{n_0} \sum_{k=-n_0}^{0} x_k \tag{44}$$

$$\mu_n = \frac{\sum_{k=1}^{n} x_k}{n + n_0} + \frac{n_0}{n + n_0} (\frac{1}{n_0}) \sum_{k=-n_0}^{0} x_k \tag{45}$$

$$n_0 = \frac{\sigma^2}{\sigma_0^2} \tag{46}$$

$$\mu_n = \frac{1}{n + n_0} [\sum_{k=1}^{n} x_k + \sum_{k=n_0}^{0} x_k] \tag{47}$$

$$= (\frac{1}{n + n_0}) \sum_{k=-n_0}^{n} x_k \tag{48}$$

$$\therefore \sigma_n^2 = \sigma^2 \sigma_0^2 n \sigma_0^2 + \sigma^2 = \frac{\sigma^2}{n + \frac{\sigma^2}{\sigma_0^2}} \tag{49}$$

$$= \frac{\sigma^2}{n + n_0} \neq \sigma^2 \tag{50}$$

Asymptotically unbiased (page 90)
PCA issue from hand out document

References

[1] R. O. Duda, P. E. Hart, and D. E. Stork. *Pattern Classification.* Wiley and Sons, 2nd edition, 2000.