Pattern Recognition and Machine Learning

James L. Crowley

ENSIMAG 3 MMIS First Semester 2010/2011

Lesson 5 17 November 2010

Estimating Parameters for a Gaussian pdf

Contents

Source:
"Pattern Recognition and Machine Learning", C. M. Bishop, Springer Verlag, 2006.

Notation

x	a variable
X	a random variable (unpredictable value)
$\vec{x}$	A vector of D variables.
$\vec{X}$	A vector of D random variables.
D	The number of dimensions for the vector $\vec{x}$ or $\vec{X}$
E	An observation. An event.
k	Class index
K	Total number of classes
C_k	The kth class.
ω_k	The statement (assertion) that E $\in C_k$
M_k	Number of examples for the class k. (think M = Mass)
M	Total number of examples. $M = \sum_{k=1}^{K} M_k$
$\{X_m^k\}$	A set of M_k examples for the class k.
$\{X_m\} = \bigcup_{k=1,K} \{X_m^k\}$	
$\{t_m\}$	A set of class labels (indicators) for the samples
$\mu = E\{X_m\}$	The Expected Value, or Average from the M samples.
$\sigma_{ML}^2 = \hat{\sigma}^2$	Estimated Variance
$\tilde{\sigma}^2$	True Variance

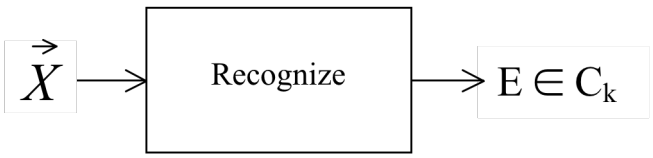
$\mathcal{N}(x;\ \mu,\ \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \quad e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ Gaussian (Normal) Density function.

The Pattern Recognition Problem

Assume that we have a sensor that produces discrete observations of the world. Each observation is an event, E. Assume that for each observation, the sensor provides a vector of D features, $\vec{X}$

Observation: $(E, \vec{X})$

Our problem is to build a box that assigns each observation to one of K classes $\{C_k\}$ labeled k=1 to K.



This problem is known as "Decision Theory". $\hat{\omega}_k = decide(E \in C_k)$

We can decompose this into two component functions $d()$ and $y(\vec{X})$:

$$\hat{\omega}_k \leftarrow d(y(\vec{X}))$$

Where $y(\vec{X})$ is a discriminant function that maps $R^D \rightarrow R^K$
d() is a decision function d(): $R^K \rightarrow \{\hat{\omega}_k\}$

Generally we choose d() to make as few mistakes as possible.
We can express this mathematically using probability theory as:

$$\hat{\omega}_k = \arg\max_{\omega_k}\{p(\omega_k \mid \vec{X})\}$$

In this case, our primary tools are Bayes Rule, that tells us:

$$p(\omega_k \mid \vec{X}) = \frac{p(\vec{X} \mid \omega_k)}{p(\vec{X})} p(\omega_k)$$

In general, $p(\vec{X})$, $p(\vec{X} \mid \omega_k)$ and $p(\omega_k)$ are estimated from a set of training data composed of M sample observations $\{\vec{X}_m\}$ labeled with an "indicator" variable $\{t_m\}$ telling the class k for each observation.

Equivalently, we can partition the training set $\{\vec{X}_m\}$ into K subsets $\{X_m^k\}$ each of which contains M_k samples.

Typically $p(\omega_k)$ is estimated as $p(\omega_k) = \frac{M_k}{M}$ although this can also be obtained from other sources.

The Gaussian density that allows us to estimate

$$p(\vec{X} \mid \omega_k) = \mathcal{N}(\vec{X} \mid \vec{\mu}_k, \Sigma_k) = \frac{1}{(2\pi)^{\frac{D}{2}} \det(\Sigma_k)^{\frac{1}{2}}} e^{-\frac{1}{2}(\vec{X}-\vec{\mu}_k)^T \Sigma_k^{-1} (\vec{X}-\vec{\mu}_k)}$$

$$p(\vec{X}) = \sum_{k=1}^{K} p(\vec{X} \mid \omega_k) = \sum_{k=1}^{K} \mathcal{N}(\vec{X} \mid \vec{\mu}_k, \Sigma_k)$$

Where the parameters $\vec{\mu}_k$ (mean) and Σ_k (covariance) for $p(\vec{X} \mid \omega_k)$, as well as $p(\omega_k)$ are estimated from the training data $\{\vec{X}_m\}$ and $\{t_m\}$.

Today we look at some of the different methods to compute this estimation.

Multivariate Gaussian Density Functions (cont'd)

More properties for Multivariate Gaussian Density Functions

Assume a feature vector $\vec{X}$ of D random variables

$$p(\vec{X}) = \mathcal{N}(\vec{X} \mid \vec{\mu}, \Sigma) = \frac{1}{(2\pi)^{\frac{D}{2}} \det(\Sigma)^{\frac{1}{2}}} e^{-\frac{1}{2}(\vec{X}-\vec{\mu})^T \Sigma^{-1} (\vec{X}-\vec{\mu})}$$

The classic method to estimate the parameters from a training set $\{\vec{X}_m\}$ as the first and second moments of the training data.

$$\vec{\mu} = E\{\vec{X}\} = \frac{1}{M}\sum_{m=1}^{M} \vec{X}_m = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \dots \\ \mu_D \end{pmatrix} = \begin{pmatrix} E\{X_1\} \\ E\{X_2\} \\ \dots \\ E\{X_D\} \end{pmatrix}$$

and

$$\Sigma = E\{(\vec{X} - E\{\vec{X}\})(\vec{X} - E\{\vec{X}\})^T\}$$

Where $$\Sigma = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \dots & \sigma_{1D}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 & \dots & \sigma_{2D}^2 \\ \dots & \dots & \dots & \dots \\ \sigma_{D1}^2 & \sigma_{D2}^2 & \dots & \sigma_{DD}^2 \end{pmatrix}$$

$$\sigma_{ij}^2 = E\{(X_i - \mu_i)(X_j - \mu_j)\} = \frac{1}{M}\sum_{m=1}^{M} (X_{im} - \mu_i)(X_{jm} - \mu_j)$$

In some cases, it is convenient to work with an inverse of the covariance:

$$\Lambda = \Sigma^{-1}$$

This is called the "precision" for the training set $\{\vec{X}_m\}$.
For example, if each observation X_m is corrupted by a sensor noise with mean 0 and covariance β, then the estimated covariance, $\hat{\Sigma}$ is

$$\hat{\Sigma}^{-1} = \Sigma^{-1} + \beta^{-1}$$

This is more conveniently expressed with precisions, as precisions add.

$$\hat{\Lambda} = \Lambda + \Lambda_B \quad \text{where } \Lambda_B = \beta^{-1}$$

The partition of a Gaussian PDF

Suppose we partition the vector $\vec{X}$ of D random variables into sub-vectors $\vec{X}_a$ and $\vec{X}_b$ of A and B components A+B=D.

$$\vec{X} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{pmatrix} = \begin{pmatrix} \vec{X}_a \\ \vec{X}_b \end{pmatrix}$$

The partition of a Gaussian random vector is composed of two Gaussian random vectors.

$$\vec{\mu} = \begin{pmatrix} \vec{\mu}_a \\ \vec{\mu}_b \end{pmatrix} \text{ and } \Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix} \text{ where } \Sigma_{ab} = \Sigma_{ab}^T$$

similarly

$$\Lambda = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix}$$

Conditional Gaussian Density

If two random vector have Gaussian statistics, then their conditional probability is Gaussian.

Suppose that $\vec{X}_a$ and $\vec{X}_b$ are both Gaussian.

$$p(\vec{X}_a \mid \vec{X}_b) = \mathcal{N}(\vec{X}_a \mid \vec{\mu}_{a|b}, \Sigma_{a|b}) = \frac{1}{(2\pi)^{\frac{A}{2}} \det(\Sigma_{a|b})^{\frac{1}{2}}} e^{-\frac{1}{2}(\vec{X}-\vec{\mu}_{a|b})^T \Sigma_{a|b}^{-1} (\vec{X}-\vec{\mu}_{a|b})}$$

where: $\vec{\mu}_{a|b} = \vec{\mu}_a + \Sigma_{ab}\Sigma_{bb}^{-1}(\vec{X}_b - \vec{\mu}_b)$

and $\Sigma_{a|b} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}$

The derivation is in Bishop pages 84-87.

Likelihood Estimation for the Gaussian Parameters

There are alternative methods to define the parameters for a Gaussian pdf.
For example, we can compute the most "likely" parameters for the data set as a maximum likelihood estimate.

Consider M sample observations $\mathbf{X} = \{\vec{X}_m\}$.

Assuming that the $\vec{X}_m$ are independent,

$$p(\vec{X}_1, \vec{X}_2 \mid \vec{\mu}, \Sigma) = p(\vec{X}_1 \mid \vec{\mu}, \Sigma) \cdot p(\vec{X}_2 \mid \vec{\mu}, \Sigma)$$

so that

$$p(\vec{X}_1, \vec{X}_2, ..., \vec{X}_M \mid \vec{\mu}, \Sigma) = \prod_{m=1}^{M} \mathcal{N}(\vec{X}_m \mid \vec{\mu}, \Sigma)$$

we define this as the Likelihood. (recall $\mathrm{X} = \{\vec{X}_m\}$.)

$$L(\vec{\mu}, \Sigma \mid \mathbf{X}) = p(\mathbf{X} \mid \vec{\mu}, \Sigma) = \prod_{m=1}^{M} \mathcal{N}(\vec{X}_m \mid \vec{\mu}, \Sigma)$$

in general is it more convenient to work with the Log-likelihood:

$$\mathcal{L}(\vec{\mu}, \Sigma \mid \mathbf{X}) = \ln\{L(\vec{\mu}, \Sigma \mid \mathbf{X})\} = \sum_{m=1}^{M} \ln\{\mathcal{N}(\vec{X}_m \mid \vec{\mu}, \Sigma)\}$$

$$\mathcal{L}(\vec{\mu}, \Sigma \mid \mathbf{X}) = \ln\{L(\vec{\mu}, \Sigma \mid \mathbf{X})\} = \sum_{m=1}^{M} \ln\{\mathcal{N}(\vec{X}_m \mid \vec{\mu}, \Sigma)\}$$

The log likelihood for X is

$$\mathcal{L}(\vec{\mu}, \Sigma \mid \mathbf{X}) = \ln\{p(\mathbf{X} \mid \vec{\mu}, \Sigma)\} = -\frac{MD}{2} \ln\{2\pi\} - \frac{M}{2} \ln\{\det(\Sigma)\} - \frac{1}{2} \sum_{m=1}^{M} (X_m - \mu)^T \Sigma^{-1} (X_m - \mu)$$

using algebra we can show that

$$\frac{\partial \mathcal{L}(\vec{\mu}, \Sigma \mid \mathbf{X})}{\partial \vec{\mu}} = \sum_{m=1}^{M} \Sigma^{-1} (X_m - \mu)$$

setting this to zero we obtain

$$\vec{\mu}_{ML} = \frac{1}{M}\sum_{m=1}^{M}\vec{X}_m$$

Similarly, but setting

$$\frac{\partial \mathcal{L}(\vec{\mu}, \Sigma \mid \mathbf{X})}{\partial \sigma_{ij}} = 0$$

we can obtain

$$\Sigma_{ML} = \frac{1}{M}\sum_{m=1}^{M}(\vec{X}_m - \mu_{ML})^T(\vec{X}_m - \mu_{ML})$$

Notice that the Maximum likelihood gives a "biased" estimate for Σ^2.

If we evaluate draw our M Samples from a normal density with

$\vec{\mu}$ and Σ

$$p(\vec{X}_m) \leftarrow \mathcal{N}(\vec{X}_m \mid \vec{\mu}, \Sigma)$$

we will discover that

$\vec{\mu}_{ML} = \vec{\mu}$ but $\Sigma_{ML} = \frac{M-1}{M}\Sigma$

The unbiased estimate would be:

$$\Sigma = \frac{1}{M-1}\sum_{m=1}^{M}(\vec{X}_m - \mu_{ML})(\vec{X}_m - \mu_{ML})^T$$

Σ_{ML} and Σ converge as M grows larger.

Sequential Estimation of Gaussian Parameters

In many on-line applications, new data must be added to the estimation as it arrives. This can be accomplished with a Bayesian approach to estimation.

In Bayesian recognition we are interested in accumulating evidence.
Each new sample X_m is evidence for $\vec{\mu}_{ML}$ and Σ_{ML}.

We can see this by reformulating the estimation sequentially, as if the data arrive in temporal sequence. The estimate after M points is:

$$\vec{\mu}_{ML}^{(M)} = \frac{1}{M}\sum_{m=1}^{M}\vec{X}_m$$

we can decompose this to

$$\vec{\mu}_{ML}^{(M)} = \frac{1}{M}\vec{X}_m + \frac{1}{M}\sum_{m=1}^{M}\vec{X}_m$$
$$\vec{\mu}_{ML}^{(M)} = \frac{1}{M}\vec{X}_m + \frac{M-1}{M}\vec{\mu}_{ML}^{(M-1)}$$
$$\vec{\mu}_{ML}^{(M)} = \vec{\mu}_{ML}^{(M-1)} + \frac{1}{M}(\vec{X}_m - \vec{\mu}_{ML}^{(M-1)})$$

We can interpret this as saying that the "influence" of the new data decreases as 1/M. Clearly, as M increases the contribution from each data point gets smaller.

Bayesian Inference for the Gaussian Parameters

Bayesian estimation considers the estimation as a problem of evidence accumulation. To keep the algebra simple, consider that case where D=1 and suppose that σ^2 is fixed.

as before, our sample set is $\mathbf{X} = \{\vec{X}_m\}$.

$$p(\mathbf{X}\mid\mu) = \prod_{m=1}^{M}\mathcal{N}(X_m\mid\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}}e^{-\frac{1}{2\sigma^2}\sum_{m=1}^{M}(X_m-\mu)^2}$$

Note that $p(\mathbf{X}|\mu)$ is NOT a pdf and does NOT sum to 1.

If we choose a prior $p(\mu) = \mathcal{N}(\mu \mid \mu_o, \sigma_o^2)$

asd

then the posterieur density is a production of two quadratics, and hence also Gaussian.

$$p(\mu \mid \mathbf{X}) = \mathcal{N}(\mu \mid \mu_M, \sigma_M^2)$$

Thus

$$p(\mu \mid \mathbf{X}) \propto p(\mathbf{X} \mid \mu) p(\mu)$$

where

$$\mu_m = \frac{\sigma^2}{M\sigma^2 + \sigma^2}\mu_o + \frac{M\sigma_o^2}{M\sigma_o^2 + \sigma^2}\mu_{ML}$$

and

$$\frac{1}{\sigma_M^2} = \frac{1}{\sigma_o^2} + \frac{M}{\sigma^2}$$

where

$$\mu_{ML} = \frac{1}{M}\sum_{m=1}^{M} X_m$$

Not that $\frac{1}{\sigma_M^2} = \frac{1}{\sigma_o^2} + \frac{M}{\sigma^2}$ is more conveniently expressed as the precision: $\lambda = 1/\sigma^2$ because precision are combined by addition.

=

$$\lambda_M = \lambda_o + M\lambda$$

Thus we can formulate:

$$p(\mu \mid \mathrm{X}) \propto \left[p(\mu) \prod_{m=1}^{M-1} p(X_m \mid \mu) \right] p(X_m \mid \mu)$$

and

$$p(\mathrm{X} \mid \lambda) = \prod_{m=1}^{M} N(X_m \mid \mu, \lambda^{-1}) \propto \lambda^{M/2} e^{\left(-\frac{\lambda}{2}\right)\sum_{m-1}^{M}(X_m - \mu)^2}$$

which is equivalent to

$$p(\mathrm{X} \mid \frac{1}{\sigma^2}) = \prod_{m=1}^{M} N(X_m \mid \mu, \sigma^2) \propto \left(\frac{1}{\sigma^2}\right)^{M/2} e^{\left(-\frac{1}{2\sigma^2}\right)\sum_{m-1}^{M}(X_m-\mu)^2}$$