Probabilistic Generative Models

Prof. Nicholas Zabaras
School of Engineering
University of Warwick
Coventry CV4 7AL
United Kingdom

Email: nzabaras@gmail.com
URL: http://www.zabaras.com/

August 1, 2014
Contents

- Probabilistic Generative Models for two Classes, Logistic Sigmoid, Models with K>2, Gaussian Class Conditionals
- Maximum Likelihood Solution, Multiclass Case
- Discrete Features
- Exponential Family, Generalization to Exponential Family Class Conditionals

- Following closely Chris Bishops’ PRML book, Chapter 4.
- K. Murphy, Machine Learning: A probabilistic Perspective, Chapter 8
Probabilistic Generative Models

- Models with linear decision boundaries arise from simple assumptions about the probabilistic distribution of the data.

- We here adopt a generative approach: we model the class-conditional densities $p(x|C_k)$, as well as the class priors $p(C_k)$, and then use these to compute posterior probabilities $p(C_k|x)$ through Bayes’ theorem.

- Let us consider 2 classes:

$$p(C_1 | x) = \frac{p(x | C_1) p(C_1)}{p(x | C_1) p(C_1) + p(x | C_2) p(C_2)} = \frac{1}{1 + e^{-a}} = \sigma(a)$$

where:

$$a = \ln \frac{p(x | C_1) p(C_1)}{p(x | C_2) p(C_2)}$$

and

$$\sigma(a) = \frac{1}{1 + e^{-a}}$$

Logistic sigmoid
Logistic Sigmoid (S-shaped)

- We can approximate the logistic sigmoid function with the scaled inverse probit function.
- The derivatives of the two curves are the same for $a=0$ (selection of $\lambda^2 = \frac{\pi}{8}$).

Logistic sigmoid

$$\sigma(a) = \frac{1}{1 + e^{-a}}$$

The logistic sigmoid function satisfies:

$$\sigma(-a) = 1 - \sigma(a)$$

The inverse of $\sigma(a)$ is the logit function:

$$a = \ln \frac{\sigma}{1 - \sigma} = \ln \frac{p(C_1|\mathbf{x})}{p(C_2|\mathbf{x})}$$

MatLab Code

$$\Phi(\lambda a) = \int_{-\infty}^{\lambda a} \mathcal{N}(\theta|0,1) \, d\theta$$

-- dashed line

Scaled inverse Probit function
The appearance of the logistic sigmoid is rather arbitrary up to this point.

\[ p(C_1 \mid x) = \frac{p(x \mid C_1)p(C_1)}{p(x \mid C_1)p(C_1) + p(x \mid C_2)p(C_2)} = \frac{1}{1 + e^{-a}} = \sigma(a) \]

This will be useful provided \( a(x) = \ln \frac{p(x \mid C_1)p(C_1)}{p(x \mid C_2)p(C_2)} \) takes a simple functional form.

We will consider problems with \( a(x) \) being a linear function of \( x \).

In this case, the posterior probability is governed by a generalized linear model.
For the case of $K > 2$ classes, we have

$$p(C_k \mid x) = \frac{p(x \mid C_k)p(C_k)}{\sum_j p(x \mid C_j)p(C_j)} = \frac{e^{a_k}}{\sum_j e^{a_j}}$$

This is known as the normalized exponential and can be regarded as a multiclass generalization of the logistic sigmoid.

Here we define:

$$a_k = \ln p(x \mid C_k)p(C_k)$$

The normalized exponential is also known as the softmax function.

It can be seen as a smoothed version of the max function because, if $a_k \geq a_j$ for all $j \neq k$, then $p(C_k \mid x) \approx 1$, and $p(C_j \mid x) \approx 0$. 

Gaussian Class Conditionals

- Assume that the class-conditional densities are Gaussians and all classes share the same covariance matrix:

\[ p(x \mid C_k) = \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2} (x - \mu_k)^T \Sigma^{-1} (x - \mu_k) \right\} \]

- We consider the case of two classes:

\[ p(C_1 \mid x) = \sigma(a), \quad a(x) = \ln \frac{p(x \mid C_1)p(C_1)}{p(x \mid C_2)p(C_2)} \]

where:

\[ a(x) = -\frac{1}{2} (x - \mu_1)^T \Sigma^{-1} (x - \mu_1) + \frac{1}{2} (x - \mu_2)^T \Sigma^{-1} (x - \mu_2) + \ln \frac{p(C_1)}{p(C_2)} = \]

\[ = \left( \mu_1^T \Sigma^{-1} \mu_1 - \mu_2^T \Sigma^{-1} \mu_2 \right)x - \frac{1}{2} \mu_1^T \Sigma^{-1} \mu_1 + \frac{1}{2} \mu_2^T \Sigma^{-1} \mu_2 + \ln \frac{p(C_1)}{p(C_2)} \]

\[ w^T \]

\[ w_0 \]
Gaussian Class Conditionals

We thus obtain a linear function of $\mathbf{x}$:

$$p(C_1 \mid \mathbf{x}) = \sigma(\mathbf{w}^T \mathbf{x} + w_0)$$

$$\mathbf{w} = \Sigma^{-1}(\mu_1 - \mu_2), \quad w_0 = -\frac{1}{2} \mu_1^T \Sigma^{-1} \mu_1 + \frac{1}{2} \mu_2^T \Sigma^{-1} \mu_2 + \ln \frac{p(C_1)}{p(C_2)}$$

Class conditional densities (left) and corresponding posteriors (right). On the right, the surface is colored using a proportion of red ink given by $p(C_1 \mid \mathbf{x})$ and a proportion of blue ink given by $p(C_2 \mid \mathbf{x}) = 1 - p(C_1 \mid \mathbf{x})$

MatLab Code
Gaussian Class Conditionals

\[ p(C_1 \mid x) = \sigma(w^T x + w_0) \]

- The decision boundaries correspond to surfaces along which the posterior probabilities \( p(C_k \mid x) \) are constant.

- They are linear in input space.

- The prior probabilities \( p(C_k) \) enter only through the bias parameter \( w_0 \) so that changing \( p(C_k) \) only leads to parallel shifts of the decision boundaries.

\[
w_0 = -\frac{1}{2} \mu_1^T \Sigma^{-1} \mu_1 + \frac{1}{2} \mu_2^T \Sigma^{-1} \mu_2 + \ln \frac{p(C_1)}{p(C_2)}\]
For the case of $K > 2$ classes, we have
\[
p(C_k | x) = \frac{p(x | C_k)p(C_k)}{\sum_j p(x | C_j)p(C_j)} = \frac{e^{a_k}}{\sum_j e^{a_j}} \quad \text{where} \quad a_k = \ln p(x | C_k)p(C_k)
\]

The decision boundaries corresponding to the minimum misclassification rate occur when the two largest posterior probabilities are equal, and so are defined by linear functions of $x$, leading to a generalized linear model.

For this case, we can see immediately that:
\[
a_k(x) = w_k^T x + w_{k0}
\]
\[
w_k = \Sigma^{-1} \mu_k, \quad w_{k0} = -\frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \ln p(C_k)
\]

$a_k(x)$ are linear functions of $x$ due to the cancellation of the quadratic terms with the shared $\Sigma$. 

*Probalistic Generative Models $K>2$*
If we relax the assumption of a shared covariance matrix and allow each $p(x|C_k)$ to have its own $\Sigma_k$, then we obtain quadratic functions of $x$ and a quadratic discriminant.

The left-hand plot shows $p(x|C_k)$, $k=1,2,3$ each a Gaussian distribution. The red & green classes have the same $\Sigma$. 

MatLab Code
The right-hand plot shows the posteriors $p(C_k|x)$, $k=1,2,3$, in which the three colors represent the posterior probabilities for the respective 3 classes (scaled).

Note that for the two classes with the same covariance, the decision boundary is linear while for the other 2 is quadratic.
Maximum Likelihood Solution

- Once we have a parametric form for $p(x|C_k)$, we can compute the parameters, together with the prior class probabilities $p(C_k)$, using MLE.

- This requires a training data set comprising of $x$ and their corresponding class labels.

- Consider the case of two classes each having a Gaussian class-conditional density with a shared covariance matrix.

- Suppose we have a data set $\{x_n, t_n\}$ where $n = 1, \ldots, N$. Here $t_n = 1$ denotes class $C_1$ and $t_n = 0$ denotes $C_2$.

- We denote the prior class probability $p(C_1) = \pi$, $p(C_2) = 1-\pi$. 

Bayesian Scientific Computing, Spring 2013 (N. Zabaras)
Maximum Likelihood Solution

- For a data point $x_n$ from class $C_1$, we have $t_n = 1$ and hence
  \[ p(x_n, C_1) = p(C_1)p(x_n | C_1) = \pi \mathcal{N}(x_n | \mu_1, \Sigma) \]

- Similarly for class $C_2$, we have $t_n = 0$
  \[ p(x_n, C_2) = p(C_2)p(x_n | C_2) = (1-\pi)\mathcal{N}(x_n | \mu_2, \Sigma) \]

- Thus the likelihood function is given by
  \[
  p(t \mid \pi, \mu_1, \mu_2, \Sigma) = \prod_{n=1}^{N} \left[ \pi \mathcal{N}(x_n | \mu_1, \Sigma) \right]^{t_n} \left[ (1-\pi) \mathcal{N}(x_n | \mu_2, \Sigma) \right]^{1-t_n}
  \]

  where $t = (t_1, \ldots, t_N)^T$.

- We maximize the log of the likelihood function.
Maximum Likelihood Solution

\[ p(t \mid \pi, \mu_1, \mu_2, \Sigma) = \prod_{n=1}^{N} \left[ \pi \mathcal{N}(x_n \mid \mu_1, \Sigma) \right]^{t_n} \left[ (1 - \pi) \mathcal{N}(x_n \mid \mu_2, \Sigma) \right]^{1-t_n} \]

- Consider first the maximization with respect to \( \pi \). The terms of interest in the log-likelihood are:

\[ \sum_{n=1}^{N} \left\{ t_n \ln \pi + (1 - t_n) \ln(1 - \pi) \right\} \]

- Taking derivative wrt \( \pi \) and setting it equal to zero:

\[ \pi = \frac{1}{N} \sum_{n=1}^{N} t_n = \frac{N_1}{N_1 + N_2} \]

where \( N_i \) the total number of data points in class \( C_i \).

- As expected, the maximum likelihood estimate for \( \pi \) is simply the fraction of points in class \( C_1 \).

- This result is easily generalized to the multiclass case (see next) where the MLE of the prior probability associated with class \( C_k \) is given by the fraction of the training set points \( N_k \).
Maximum Likelihood Solution: Multiclass

\[ p(\{\phi_n, t_n\} | \{\pi_k\}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \left[ \pi_k p(\phi_n | C_k) \right]^{t_{nk}} \]

- For the multiclass case, the log likelihood takes the form:

\[ \ln p(\{\phi_n, t_n\} | \{\pi_k\}) = \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \left( \ln \pi_k + \ln p(\phi_n | C_k) \right), \quad \sum_{k=1}^{K} \pi_k = 1 \]

- To maximize the log likelihood, we consider:

\[ \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \left( \ln \pi_k + \ln p(\phi_n | C_k) \right) + \lambda \left( \sum_{k=1}^{K} \pi_k - 1 \right) \]

- Taking the derivative wrt \( \pi_k \) we arrive at the same result as for two classes:

\[ \sum_{n=1}^{N} \frac{t_{nk}}{\pi_k} + \lambda = 0 \Rightarrow \pi_k \lambda = -\sum_{n=1}^{N} t_{nk} = -N_k \Rightarrow \left\{ \begin{array}{l} \sum_{k=1}^{K} \pi_k \lambda = -N \\ \pi_k \lambda = -N_k \end{array} \right. \Rightarrow \left\{ \begin{array}{l} \lambda = -N \\ \pi_k \lambda = -N_k \Rightarrow \pi_k = \frac{N_k}{N} \end{array} \right. \]
Maximum Likelihood Solution

\[ p(t | \pi, \mu_1, \mu_2, \Sigma) = \prod_{n=1}^{N} \left[ \pi \mathcal{N}(x_n | \mu_1, \Sigma) \right]^{t_n} \left[ (1 - \pi) \mathcal{N}(x_n | \mu_2, \Sigma) \right]^{1-t_n} \]

- Consider now the maximization with respect to \( \mu_1 \). The terms of interest in the log-likelihood are:

\[ \sum_{n=1}^{N} t_n \ln \mathcal{N}(x_n | \mu_1, \Sigma) = -\frac{1}{2} \sum_{n=1}^{N} t_n (x_n - \mu_1)^T \Sigma^{-1} (x_n - \mu_1) + \text{const.} \]

- Taking derivative wrt \( \mu_1 \) and setting it equal to zero:

\[ \mu_1 = \frac{1}{N_1} \sum_{n=1}^{N} t_n x_n, \text{ and similarly } \mu_2 = \frac{1}{N_2} \sum_{n=1}^{N} (1-t_n) x_n \]

- Thus the MLE for \( \mu_i \) is simply the mean of the fraction of points in class \( \mathcal{C}_i \).

- This result is easily generalized to the multiclass case. The MLE of the prior probability associated with class \( \mathcal{C}_k \) is given by the fraction of the training set points \( N_k \).
Maximum Likelihood Solution

\[ p(t \mid \pi, \mu_1, \mu_2, \Sigma) = \prod_{n=1}^{N} \left[ \pi \mathcal{N}(x_n \mid \mu_1, \Sigma) \right]^{t_n} \left[ (1 - \pi) \mathcal{N}(x_n \mid \mu_2, \Sigma) \right]^{1-t_n} \]

Consider finally the maximization with respect to \( \Sigma \). The terms of interest in the log-likelihood are:

\[
-\frac{1}{2} \sum_{n=1}^{N} t_n \ln |\Sigma| - \frac{1}{2} \sum_{n=1}^{N} t_n (x_n - \mu_1)^T \Sigma^{-1} (x_n - \mu_1) \\
-\frac{1}{2} \sum_{n=1}^{N} (1-t_n) \ln |\Sigma| - \frac{1}{2} \sum_{n=1}^{N} (1-t_n) (x_n - \mu_2)^T \Sigma^{-1} (x_n - \mu_2) \\
= -\frac{N}{2} \ln |\Sigma| - \frac{N}{2} Tr \left( \Sigma^{-1} S \right), \text{ where } S = \frac{N_1}{N} S_1 + \frac{N_2}{N} S_2 \\
S_1 = \frac{1}{N_1} \sum_{n \in C_1} (x_n - \mu_1)(x_n - \mu_1)^T, \ S_2 = \frac{1}{N_2} \sum_{n \in C_2} (x_n - \mu_2)(x_n - \mu_2)^T
\]

Using the result for the MLE for a Gaussian distribution, we obtain \( \Sigma = S \). This result is a weighted average of the covariance matrices associated with each class separately.
These results are easily extended to the $K$ class problem to obtain the corresponding MLE solutions for the parameters (for Gaussian class-conditionals with a shared covariance matrix).

As we have seen before, MLE is not robust to outliers.

- Fitting Gaussian distributions to the classes is not robust to outliers, because the MLE of a Gaussian is not robust.
Here we consider the case of discrete feature values \( x_i \). Consider binary feature values \( x_i \in \{0, 1\}, i=1,\ldots,D \).

For \( D \) inputs, a general distribution corresponds to a table of \( 2^D \) numbers for each class, containing \( 2^D - 1 \) independent variables (due to the summation constraint).

This grows exponentially with the number \( D \) of features.

Consider the Naive Bayes Assumption: the feature values are treated as independent, conditioned on the class \( C_k \).

Thus we have class-conditional distributions of the form

\[
p(x \mid C_k) = \prod_{i=1}^{D} \mu_{ki}^{x_i} (1 - \mu_{ki})^{1-x_i}
\]
Discrete Features

\[ p(x \mid C_k) = \prod_{i=1}^{D} \mu_{ki}^{x_i} (1 - \mu_{ki})^{1-x_i} \]

- There are \( D \) independent parameters for each class.
- Here \( \mu_{ki} \) is the probability the \( i^{th} \) feature takes the value 1.
- Substituting into \( a_k = \ln p(x \mid C_k) p(C_k) \) gives:

\[
a_k(x) = \sum_{i=1}^{D} \left\{ x_i \ln \mu_{ki} + (1-x_i) \ln (1 - \mu_{ki}) \right\} + \ln p(C_k)
\]

- These are linear functions of the input values \( x_i \).
- For \( K = 2 \) classes, we can alternatively consider the logistic sigmoid formulation. Analogous results are obtained for discrete variables each of which can take \( M > 2 \) states.
We have seen that for both Gaussian and discrete inputs, the posterior class probabilities are given by generalized linear models with logistic sigmoid ($K=2$) or softmax ($K \geq 2$ classes) activation functions.

This result is general for the class-conditional densities $p(x|C_k)$ that are members of the exponential family.

For the exponential family, the distribution of $x$ can be written in the form

$$p(x | \lambda_k) = h(x) g(\lambda_k) \exp \left\{ \lambda_k^T u(x) \right\}$$

Here we consider the case $u(x) = x$. 
Exponential Family

\[ p(x \mid \lambda_k) = h(x) g(\lambda_k) \exp \left\{ \lambda_k^T u(x) \right\} \]

- With \( u(x) = x \) and by introducing a scaling parameter \( s \) as \( p(x \mid s) = (1/s)f(x/s) \), we obtain the restricted set of exponential family class-conditional densities of the form:

\[ p(x \mid \lambda_k, s) = \frac{1}{s} h\left(\frac{1}{s} x\right) g(\lambda_k) \exp \left\{ \frac{1}{s} \lambda_k^T x \right\} \]

- Here each class has its own parameter \( \lambda_k \) but all classes share the same scale parameter \( s \).

- For the 2 class problem, we substitute this expression for the class-conditional densities into

\[ a(x) = \ln \frac{p(x \mid C_1) p(C_1)}{p(x \mid C_2) p(C_2)} \]
Exponential Family

\[ p(x \mid \lambda_k, s) = \frac{1}{s} h\left(\frac{1}{s} x\right) g(\lambda_k) \exp \left\{ \frac{1}{s} \lambda_k^T x \right\} \]

\[ a(x) = \ln \frac{p(x \mid C_1) p(C_1)}{p(x \mid C_2) p(C_2)} \]

- The posterior class probability is again given by a logistic sigmoid acting on a linear function \( a(x) \)

\[ p(C_1 \mid x) = \sigma(a(x)) \]

which is given by

\[ a(x) = \frac{1}{s} \left( \lambda_1 - \lambda_2 \right)^T x + \ln g(\lambda_1) - \ln g(\lambda_2) + \ln p(C_1) - \ln p(C_2) \]

- Similarly for the K-class problem, we substitute the class-conditionals directly in \( a_k(x) = \ln p(x \mid C_k) p(C_k) \) to obtain:

\[ a(x) = \frac{1}{s} \lambda_k^T x + \ln g(\lambda_k) + \ln p(C_k) \]

which is linear in \( x \).