Statistical and Learning Techniques in Computer Vision Lecture 2: Maximum Likelihood and Bayesian Estimation

Jens Rittscher and Chuck Stewart

1 Motivation and Problem

- In Lecture 1 we briefly saw how histograms could be used to model the probability of pixel intensity values:

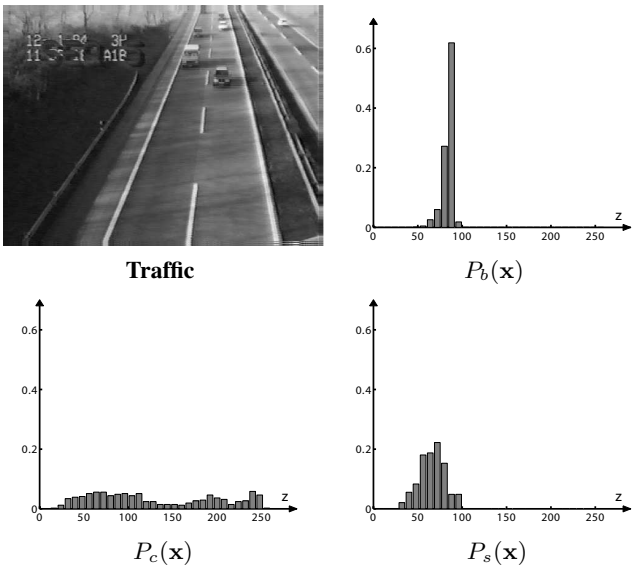

Figure 1: **Pixels as random variables.** *The image shows a sample frame of a surveillance camera installed to monitor traffic. In order to build a background model of the scene the set of pixels was devided into three different sets: background, cars, and shadow. The three graphs show the corresponding histograms of the grey values of pixel locations from the different sets. Notice that the distribution of the background pixels $P_b(\mathbf{x})$ is sharply peaked. The distribution of greyvalues of pixels $P_c(\mathbf{x})$ that correspond to vehicles is, on the other hand, almost uniform.*

- Histogram-based modeling is just one (simple) solution to a question of primary interest in this course: how to model probability density functions.
  - Many learning problems may be cast as distribution modeling problems.
- In the next three lectures we will discuss density modeling techniques, including techniques more sophisticated than histograms
  - As we will see, there are several limitations to the use of histograms.

- Here is a more careful, though abstract, statement of our problem:

  > Given a finite sample of data points $\mathcal{X} = \{\mathbf{x}^i\}$ all drawn from the same underlying density $p(x)$, determine an accurate model of $p(x)$.

  - We will generally be interested in representation of continuous densities.
  - For now, the superscript on the data refers to the index of the (vector-valued) sample, whereas subscripts will indicate the index of a component within a particular sample.

2 Modeling approaches

- *Parametric models:* the probability density function has a specific functional form that depends on a vector of parameters, $\boldsymbol{\theta}$. We write this as the conditional density

$$p(\mathbf{x}|\boldsymbol{\theta}) \tag{1}$$

  This is the *probability of* $\mathbf{x}$ *given parameter vector* $\boldsymbol{\theta}$.

  - The normal distribution is an example of a parametric model.

- *Non-parametric:* The form of the density is entirely determined by the data without any model. The histograms in figure 1 are an example of such a non-parametric representation.

- *Semi-parametric:* The combination of parametric and non-parametric models allows a broad class of functional forms in which the number of parameters can be increased in a systematic way to build more flexible models. The total number of parameters can be varied independent of the size of the data set. Gaussian mixture models are an example of such a class of distributions.

In general, the difficulty of generating accurate probability models increases dramatically in high dimensions.

3 Lectures 2 through 4

Today's lecture and Lectures 3 and 4 cover the above three types of modeling approaches, using them to study a number of different techniques that will be important throughout the semester.

- Today: parametric models, maximum likelihood estimation and Bayesian estimation

- Lecture 3: kernel based methods, plus an application in registration based on mutual information.

- Lecture 4: semi-parametric models, mixture models and the expectation maximization algorithm.

4 Maximum Likelihood Estimation

- Recall that we are given samples $\mathcal{X} = \{\mathbf{x}^i\}$ and we want to estimate the parameters $\boldsymbol{\theta}$. The density model is
$$p(\mathbf{x}|\boldsymbol{\theta}). \tag{2}$$

- For a given set of parameters $\boldsymbol{\theta}$, the probability of an individual $\mathbf{x}^i$ is
$$p(\mathbf{x}^i|\boldsymbol{\theta}). \tag{3}$$

- Making the reasonable assumption that the points are independent of each other (and of course follow the same distribution), the probability of obtaining $\mathcal{X}$ is
$$p(\mathcal{X}|\boldsymbol{\theta}) = \prod_{i=1}^{N} p(\mathbf{x}^i|\boldsymbol{\theta}). \tag{4}$$

- This is the "likelihood function" for a particular parameter vector $\boldsymbol{\theta}$, written as
$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} p(\mathbf{x}^i|\boldsymbol{\theta}). \tag{5}$$

  Our goal is to find the value of $\boldsymbol{\theta}$ that maximizes this likelihood.

- Because the log is a monotonic function, we can obtain the same result by minimizing the negative log of $L(\boldsymbol{\theta})$:
$$-\ln L(\boldsymbol{\theta}) = l(\boldsymbol{\theta}) = \sum_{i=1}^{N} \ln p(\mathbf{x}^i|\boldsymbol{\theta}). \tag{6}$$

  This form is often simpler and therefore the one usually used.

- Given the resulting estimate $\hat{\boldsymbol{\theta}}$, the desired density is
$$p(\mathbf{x}|\hat{\boldsymbol{\theta}}. \tag{7}$$

5 Two Examples

In class we will work through two examples:

1. $p(\mathbf{x}|\boldsymbol{\theta})$ is the univariate Gaussian distribution.
   - In this case $\boldsymbol{\theta} = (\mu, \sigma)^\top$.
   - The result will be the maximum-likelihoood estimates of the mean and variance of the Gaussian distribution.
   - We will show that the resulting estimate is (slightly) biased.

2. $p(\mathbf{x}|\boldsymbol{\theta})$ is the error of a point from a regression plane.

   - The form of the regression plane is the explicit function

$$\begin{aligned} y &= a_0 + a_1 x_1 + \cdots a_k x_k \\ &= (a_0, \ldots, a_k) \begin{pmatrix} 1 \\ \tilde{\mathbf{x}} \end{pmatrix} \\ &= \mathbf{a}^\top \begin{pmatrix} 1 \\ \tilde{\mathbf{x}} \end{pmatrix} \end{aligned} \tag{8}$$

   - The data points are $\mathbf{x}^\top = (\tilde{\mathbf{x}}^\top, y)$, where y is a measurement and $\tilde{\mathbf{x}}$ is the location at which the measurement is made.
   - If we assume the measurement errors are normally distributed from the line and are independent with unknown variance σ^2, then the parameter vector to be estimated is

$$\boldsymbol{\theta}^\top = (\mathbf{a}^\top, \sigma^2) \tag{9}$$

   - Here there is no mean parameter, μ. Instead, the plane plays the role of the mean vector.
   - We will show in class that maximum likelihood estimation leads to the usual least-squares plane parameter estimate.

6 Bayesian Density Estimation — Overview

We will consider this in two parts

1. Computing the probability

$$p(\boldsymbol{\theta}|\mathcal{X}) \tag{10}$$

   which is the conditional ("posterior") probability density function of the parameter vector given data $\mathcal{X}$.

   - Formally, this is very different from maximum likelihood estimation, which is only interested in a particular (most likely) set of estimated parameters.
   - Recall that the Maximum Likelihood estimate was determined by maximizing $p(\mathcal{X}|\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$.
   - The important new information that is now being considered is the prior distribution $p(\boldsymbol{\theta})$.

2. Use the estimated probability $p(\boldsymbol{\theta}|\mathcal{X})$ to estimate the probability $p(\mathbf{x}|\mathcal{X})$, i.e. the density of the variable $\mathbf{x}$ given the samples $\mathcal{X}$.

   - We do not write $p(\mathbf{x}|\boldsymbol{\theta})$ as with maximum likelihood estimation because we will eliminate $\boldsymbol{\theta}$ through integration! This allows us to consider all possible values of the parameter vector $\boldsymbol{\theta}$, in a true Bayesian style,
   - This is used when you want to classify a new measurement based on the probability of obtaining it given previous measurements.

7 Bayesian Density Estimation — Part 1

Estimating $p(\boldsymbol{\theta}|\mathcal{X})$, the posterior distribution:

- Using Bayes' theorem,

$$p(\boldsymbol{\theta}|\mathcal{X}) = \frac{p(\mathcal{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{X})}. \tag{11}$$

- Notice that the denominator does not involve $\boldsymbol{\theta}$. Therefore we ignore it (for now) and later normalize the resulting function to make it a density. Dropping the denominator gives

$$p(\boldsymbol{\theta}|\mathcal{X}) \propto p(\mathcal{X}|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{12}$$

- The first factor is the conditional density from Equation 4 that we used to obtain the maximum likelihood estimate.
- The second factor is the prior probability.
  - It represents what is known in advance about the parameter vector.
- The result, $p(\boldsymbol{\theta}|\mathcal{X})$, represents what is known about $\boldsymbol{\theta}$ after analyzing the data set $\mathcal{X}$. This is known as the *posterior* density.
- Closed forms for $p(\boldsymbol{\theta}|\mathcal{X})$ may only be obtained in relatively simple circumstances. In general, numerical methods are needed.
- The *maximum a posteriori estimate* MAP of the parameter $\boldsymbol{\theta}$ is defined as

$$\hat{\boldsymbol{\theta}} = \max_{\boldsymbol{\theta}} \, p(\boldsymbol{\theta}|\mathcal{X}) \ . \tag{13}$$

- As a simple example, to be worked in class, we will estimate the posterior distribution of the mean μ of a univariate distribution, assuming
  - the samples are normally distributed with the unknown mean μ and a known variance σ^2, and
  - the prior of μ is normally distributed itself with mean μ_0 and variance σ_0^2.

8 Bayesian Density Estimation — Part 2

Estimating $p(\mathbf{x}|\mathcal{X})$ from $p(\boldsymbol{\theta}|\mathcal{X})$

- We have a density that ranges over all $\boldsymbol{\theta}$. The problem is to obtain a density of $\mathbf{x}$.
- This is done using marginalization:

$$p(\mathbf{x}|\mathcal{X}) = \int p(\mathbf{x}, \boldsymbol{\theta}|\mathcal{X}) d\boldsymbol{\theta}. \tag{14}$$

  On the right is the joint density of $\mathbf{x}$ and $\boldsymbol{\theta}$ (given $\mathcal{X}$) and the $\boldsymbol{\theta}$ term is integrated out (in the expression).

- Next we use the rules of conditional densities and the properties of the resulting densities to write:

$$
\begin{aligned}
p(\mathbf{x}|\mathcal{X}) &= \int p(\mathbf{x}, \boldsymbol{\theta}|\mathcal{X}) d\boldsymbol{\theta} \\
&= \int p(\mathbf{x}|\boldsymbol{\theta}, \mathcal{X}) p(\boldsymbol{\theta}|\mathcal{X}) d\boldsymbol{\theta} \\
&= \int p(\mathbf{x}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{X}) d\boldsymbol{\theta} \qquad (15)
\end{aligned}
$$

  In the final integral, the first factor is just the functional form of the density model and it does not depend on $\mathcal{X}$, while the second factor is the posterior probability estimated in part 1.

- This integral is generally quite complex. We will proceed with our example from above in class.

9 Maximum Likelihood vs. Bayesian

- The maximum likelihood estimate (MLE) is a single estimated parameter vector, and is used as the parameter vector in our parametric density modeling problem to produce the final density.
- Bayesian estimation produces a posterior distribution (12) and this distribution (after proper normalization) is then combined with the parametric density model through multiplication and integration to produce our final density.
- In many cases we are only interested in the parameter estimate.
  - In this case the Bayesian posterior is maximized to produce the desired estimate. This is called the maximum a posteriori (MAP) estimate.
  - The MLE and MAP estimates are close when the data overwhelms the prior and exactly the same when the prior used is "non-informative".
- Note that in all these cases we always consider one particular model. Later on we will us a similar analysis in order to perform model selection.

10 Further reading

- The presentation here is mostly based on Duda, Hart and Stork [DHS01, Chapter 3].
- A different, but nice summary of the material relevant to probability density estimation can be found in Bishop [Bis06, Chapter 1].

References

[Bis06] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[DHS01] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. John Wiley and Sons, 2001.