Kernel-based density estimation

Nuno Vasconcelos

ECE Department, UCSD
Announcement

- last class, November 30, we will have “Cheetah Day”

- what:
  - 5 teams, average of 4 people
  - each team will write a report on the 5 cheetah problems
  - each team will give a presentation on one of the problems

- why:
  - to make sure that we get the “big picture” out of all this work
  - presenting is always good practice
  - because I am such a good person...
Announcement

► how much:
  • 10% of the final grade (5% report, 5% presentation)

► what to talk about:
  • report: comparative analysis of all solutions of the problem (8 page)
  • as if you were writing a conference paper
  • presentation: will be on one single problem
    • review what solution was
    • what did this problem taught us about learning?
    • what “tricks” did we learn solving it?
    • how well did this solution do compared to others?
Announcement

details:

• get together and form groups
• let me know what they are by November 18 (email is fine)
• I will randomly assign the problem on which each group has to be expert
• prepare a talk for 15min (7 or 8 slides)
• feel free to use my solutions, your results, create new results, whatever...
Plan for today

we have talked a lot about the BDR and methods based on density estimation

practical densities are not well approximated by simple probability models

last lectures: one alternative is to reduce the dimensionality of the feature space
  • PCA, LDA
  • Rayleigh quotient maximization

today: what can we do if need a complicated space
  • use the BDR
  • use better probability density models!
Non-parametric density estimates

Given iid training set $\mathcal{D} = \{x_1, \ldots, x_n\}$, the goal is to estimate $P_X(x)$.

Consider a region $\mathcal{R}$, and define

$$P = P_X[x \in \mathcal{R}] = \int_{\mathcal{R}} P_X(x) dx.$$  

and define

$$K = \# \{x_i \in \mathcal{D}|x_i \in \mathcal{R}\}.$$  

This is a binomial distribution of parameter $P$

$$P_K(k) = \mathcal{B}(n, P)$$

$$= \binom{n}{k} P^k (1 - P)^{n-k}$$
Binomial random variable

- ML estimate of $P$

$$\hat{P} = \frac{k}{n}.$$  

and statistics

$$E[\hat{P}] = \frac{1}{n} E[k] = \frac{1}{n} nP = P$$

$$\text{var}[\hat{P}] = \frac{1}{n^2} \text{var}[k] = \frac{P(1-P)}{n}.$$  

- Note that $\text{var}[\hat{P}] \leq 1/4n$ goes to zero very quickly, i.e.

$$\hat{P} \rightarrow P.$$  

<table>
<thead>
<tr>
<th>$N$</th>
<th>10</th>
<th>100</th>
<th>1,000</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Var[$P$]</td>
<td>0.025</td>
<td>0.0025</td>
<td>0.00025</td>
<td></td>
</tr>
</tbody>
</table>
Histogram

- this means that \( k/n \) is a very good estimate of \( P \)
- on the other hand, from the mean value theorem, if \( P_X(x) \) is continuous \( \exists \varepsilon \in \mathcal{R} \) such that

\[
P = \int_{\mathcal{R}} P_X(x) \, dx = P_X(\varepsilon) \int_{\mathcal{R}} dx = P_X(\varepsilon) V(\mathcal{R}).
\]

- this is easiest to see in 1D
  - can always find a box such that the integral of the function is equal to that of the box
  - since \( P_X(x) \) is continuous there must be a \( \varepsilon \) such that \( P_X(\varepsilon) \) is the box height
Histogram

▸ hence

\[ P_X(\epsilon) = \frac{P}{V(\mathcal{R})} \approx \frac{\hat{P}}{V(\mathcal{R})} = \frac{k}{nV(\mathcal{R})} \]

▸ using continuity of \( P_X(x) \) again and assuming \( R \) is small

\[ P_X(x) \approx \frac{k}{nV(\mathcal{R})}, \ \forall x \in V(\mathcal{R}) \]

▸ this is the histogram

▸ it is the simplest possible non-parametric estimator

▸ can be generalized into kernel-based density estimator
Kernel density estimates

- Assume \( \mathcal{R} \) is the \( d \)-dimensional cube of side \( h \)

\[
V = h^d
\]

and define indicator function of the unit hypercube

\[
\phi(u) = \begin{cases} 
1, & \text{if } |u_i| < 1/2 \\
0, & \text{otherwise.}
\end{cases}
\]

Hence

\[
\phi \left( \frac{x - x_i}{h} \right) = 1
\]

iff \( x_i \in \) hypercube of volume \( V \) centered at \( x \).

- The number of sample points in the hypercube is

\[
k_n = \sum_{i=1}^{n} \phi \left( \frac{x - x_i}{h} \right)
\]
Kernel density estimates

this means that the histogram can be written as

$$P_X(x) = \frac{1}{nh^d} \sum_{i=1}^{n} \phi \left( \frac{x - x_i}{h} \right)$$

which is equivalent to:

- “put a box around $X$ for each $X_i$ that lands on the hypercube”
- can be seen as a very crude form of interpolation
- better interpolation if contribution of $X_i$ decreases with distance to $X$

consider other windows $\phi(x)$
what sort of functions are valid windows?

note that $P_X(x)$ is a pdf if and only if

$$P_X(x) \geq 0, \forall x \text{ and } \int P_X(x)dx = 1$$

since

$$\int P_X(x)dx = \frac{1}{nh^d} \sum_{i=1}^{n} \int \phi \left( \frac{x-x_i}{h} \right) dx$$

$$= \frac{1}{nh^d} \sum_{i=1}^{n} \int \phi (y) h^d dy$$

$$= \frac{1}{n} \sum_{i=1}^{n} \int \phi (y) dy$$

these conditions hold if $\phi(x)$ is itself a pdf

$$\phi(x) \geq 0, \forall x \text{ and } \int \phi(x)dx = 1$$
Gaussian kernel

- probably the most popular in practice

\[ \phi(x) = \frac{1}{\sqrt{2\pi}^d} e^{-\frac{1}{2}x^T \Sigma x} \]

- note that \( P_X(x) \) can also be seen as a sum of pdfs centered on the \( X_i \) when \( \phi(x) \) is symmetric in \( X \) and \( X_i \)

\[ P_X(x) = \frac{1}{nh^d} \sum_{i=1}^{n} \phi\left( \frac{x - x_i}{h} \right) \]
Gaussian kernel

- **Gaussian case** can be interpreted as
  - sum of \( n \) Gaussians centered at the \( X_i \) with covariance \( h I \)
  - more generally, we can have a full covariance

\[
P_X(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} e^{-\frac{1}{2}(x-x_i)^T\Sigma^{-1}(x-x_i)}
\]

- sum of \( n \) Gaussians centered at the \( X_i \) with covariance \( \Sigma \)
- Gaussian kernel density estimate: “approximate the pdf of \( X \) with a sum of Gaussian bumps”

![Gaussian kernel density estimate graph](image)
Kernel bandwidth

back to the generic model

\[ P_X(x) = \frac{1}{nh^d} \sum_{i=1}^{n} \phi \left( \frac{x - x_i}{h} \right) \]

what is the role of \( h \) (bandwidth parameter)?

defining

\[ \delta(x) = \frac{1}{h^d} \phi \left( \frac{x}{h} \right) \]

we can write

\[ P_X(x) = \frac{1}{n} \sum_{i=1}^{n} \delta(x - x_i) \]

i.e. a sum of translated replicas of \( \delta(x) \)
Kernel bandwidth

- $h$ has two roles:
  1. rescale the $x$-axis
  2. rescale the amplitude of $\delta(x)$

- this implies that for large $h$:
  1. $\delta(x)$ has low amplitude
  2. iso-contours of $h$ are quite distant from zero
     ($x$ large before $\phi(x/h)$ changes significantly from $\phi(0)$)

\[
\delta(x) = \frac{1}{h^d} \phi \left( \frac{x}{h} \right)
\]
Kernel bandwidth

for small $h$:

1. $\delta(x)$ has large amplitude

2. iso-contours of $h$ are quite close to zero
   ($x$ small before $\phi(x/h)$ changes significantly from $\phi(0)$)

$\delta(x) = \frac{1}{h^d} \phi \left( \frac{x}{h} \right)$

what is the impact of this on the quality of the density estimates?
Kernel bandwidth

- it controls the smoothness of the estimate
  - as $h$ goes to zero we have a sum of delta functions (very “spiky” approximation)
  - as $h$ goes to infinity we have a sum of constant functions (approximation by a constant)
  - in between we get approximations that are gradually more smooth
Kernel bandwidth

- why does this matter?
- when the density estimates are plugged into the BDR
- smoothness of estimates determines the smoothness of the boundaries

![Graph showing less smooth and more smooth boundaries]

- this affects the probability of error!
Convergence

- since $P_x(x)$ depends on the sample points $X_i$, it is a random variable
- as we add more points, the estimate should get “better”
- the question is then whether the estimate ever converges
- this is no different than parameter estimation
- as before, we talk about convergence in probability
- $\hat{P}_X(x)$ converges to $P_X(x)$ if

$$\lim_{n \to \infty} E_{x_1, \ldots, x_n}[\hat{P}_X(x)] = \hat{P}_X(x)$$

$$\lim_{n \to \infty} var_{x_1, \ldots, x_n}[\hat{P}_X(x)] = 0$$
Convergence of the mean

from the linearity of $P_X(x)$ on the kernels

\[
E_{X_1, \ldots, X_n} [\hat{P}_X(x)] = \\
= \frac{1}{nh^d} \sum_{i=1}^{n} E_{X_i} \left[ \phi \left( \frac{x - x_i}{h} \right) \right] \\
= \frac{1}{n} \sum_{i=1}^{n} \int \frac{1}{h^d} \phi \left( \frac{x - v}{h} \right) P_X(v) dv \\
= \int \frac{1}{h^d} \phi \left( \frac{x - v}{h} \right) P_X(v) dv \\
= \int \delta(x - v) P_X(v) dv
\]
Convergence of the mean

hence

\[ E_{X_1, \ldots, X_n}[\hat{P}_X(x)] = \int \delta(x - v) P_X(v) dv \]

this is the convolution of \( P_X(x) \) with \( \delta(x) \)

it is a blurred version (“low-pass filtered”) unless \( h = 0 \)

in this case \( \delta(x-v) \) converges to the Dirac delta and so

\[ \lim_{h \to 0} E_{X_1, \ldots, X_n}[\hat{P}_X(x)] = P_X(x) \]
Convergence of the variance

since the $X_i$ are iid

$$\text{var}_{X_1,\ldots,X_n}[\hat{P}_X(x)] =$$

$$\sum_{i=1}^{n} \text{var}_{X_i} \left[ \frac{1}{nh^d} \phi \left( \frac{x - x_i}{h} \right) \right]$$

$$\leq n E_X \left[ \frac{1}{n^2 h^{2d}} \phi^2 \left( \frac{x - x_i}{h} \right) \right]$$

$$= \frac{1}{nh^d} \int \frac{1}{h^d} \phi^2 \left( \frac{x - v}{h} \right) P_X(v) dv$$

$$\leq \frac{1}{nh^d} \sup_{x} \left[ \phi \left( \frac{x}{h} \right) \right] \int \frac{1}{h^d} \phi \left( \frac{x - v}{h} \right) P_X(v) dv$$

$$= \frac{1}{nh^d} \sup_{x} \left[ \phi \left( \frac{x}{h} \right) \right] E_{X_1,\ldots,X_n}[\hat{P}_X(x)]$$
Convergence

In summary

\[ E_{X_1, \ldots, X_n}[\hat{P}_X(x)] = \delta(x) \odot P_X(x) \]

\[ \text{var}_{X_1, \ldots, X_n}[\hat{P}_X(x)] = \leq \frac{1}{nh^d} \sup \left[ \phi \left( \frac{x}{h} \right) \right] E_{X_1, \ldots, X_n}[\hat{P}_X(x)] \]

This means that:

- to obtain small bias we need \( h \sim 0 \)
- to obtain small variance we need \( h \) infinite
Convergence

- Intuitively makes sense
  - \( h \sim 0 \) means a Dirac around each point
  - Can approximate any function arbitrarily well
  - There is no bias
  - But if we get a different sample, the estimate is likely to be very different
  - There is large variance
  - As before, variance can be decreased by getting a larger sample
  - But, for fixed \( n \), smaller \( h \) always means greater variability

Example: fit to \( N(0,1) \) using \( h = h_1/n^{1/2} \)
Example

- **small \( h \):** spiky
  - need a lot of points to converge (variance)

- **large \( h \):** approximate \( N(0, I) \) with a sum of Gaussians of larger covariance
  - will never have zero error (bias)
Optimal bandwidth

we would like

- \( h \sim 0 \) to guarantee zero bias
- zero variance as \( n \) goes to infinity

solution:

- make \( h \) a function of \( n \) that goes to zero
- since variance is \( O(1/nh^d) \) this is fine if \( nh^d \) goes to infinity

hence, we need

\[
\lim_{n \to \infty} h(n) = 0 \quad \text{and} \quad \lim_{n \to \infty} nh(n) = \infty
\]

optimal sequences exist, e.g.

\[
h(n) = \frac{k}{\sqrt{n}} \quad \text{or} \quad h(n) = \frac{k}{\log n}
\]
Optimal bandwidth

- in practice this has limitations
  - does not say anything about the finite data case (the one we care about)
  - still have to find the best k

- usually we end up using trial and error or techniques like cross-validation
Cross-validation

**basic idea:**
- leave some data out of your training set (cross validation set)
- train with different parameters
- evaluate performance on cross validation set
- pick best parameter configuration
Leave-one-out cross-validation

- many variations

- leave-one-out CV:
  - compute $n$ estimators of $P_X(x)$ by leaving one $X_i$ out at a time
  - for each $P_X(x)$ evaluate $P_X(X_i)$ on the point that was left out
  - pick $P_X(x)$ that maximizes this likelihood
Any Questions?