Density Estimation (Lecture 2.1)

Alexander J. Smola
Alex.Smola@anu.edu.au

Machine Learning Program
National ICT Australia
RSISE, The Australian National University

This Week's Topics

- Probability Theory Basics
- Maximum Likelihood
- Priors
- Exponential Family

Probability

Basic Idea

We have events in a space of possible outcomes. Then $\mathrm{P}(X)$ tells us how likely is that an event $\mathbf{x} \in X$ will occur.

Basic Axioms

- $\Pr(X) \in [0, 1]$ for all $X \subseteq \mathfrak{X}$
- $\Pr(\mathfrak{X}) = 1$
- $\Pr\left(\cup_i X_i\right) = \sum_i \Pr(X_i)$ if $X_i \cap X_j = \emptyset$ for all $i \neq j$

Simple Corollary

$$\Pr(X \cup Y) = \Pr(X) + \Pr(Y) - \Pr(X \cap Y)$$

Multiple Variables

Two Sets

Assume that $\mathcal{X}$ and $\mathcal{Y}$ are a probability measure on the product space of $\mathcal{X}$ and $\mathcal{Y}$. Consider the space of events $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$.

Independence

If $\mathbf{x}$ and $\mathbf{y}$ are independent, then for all $X \subset \mathcal{X}$ and $Y \subset \mathcal{Y}$

$$\Pr(X, Y) = \Pr(X) \cdot \Pr(Y).$$

Dependence and Conditional Probability

Typically, knowing $\mathbf{x}$ will tell us something about $\mathbf{y}$ (think regression or classification). We have

$$\Pr(Y|X) \Pr(X) = \Pr(Y, X) = \Pr(X|Y) \Pr(Y).$$

- Hence $\Pr(Y, X) \leq \min(\Pr(X), \Pr(Y))$.
- Bayes Rule $\Pr(X|Y) = \frac{\Pr(Y|X) \Pr(X)}{\Pr(Y)}$.

Examples

How likely is it to have AIDS if the test says so?

- Assume that roughly 0.1% of the population is infected.
- The AIDS test reports positive for all infections.
- The AIDS test reports positive for 1% healthy people.

We use Bayes rule to infer $\Pr(\text{AIDS}|\text{test positive})$ via

$$\frac{\Pr(Y|X)\Pr(X)}{\Pr(Y)} = \frac{\Pr(Y|X)\Pr(X)}{\Pr(Y|X)\Pr(X) + \Pr(Y|\mathfrak{X}\backslash X)\Pr(\mathfrak{X}\backslash X)}$$
$$= \frac{1 \cdot 0.001}{1 \cdot 0.001 + 0.01 \cdot 0.999} = 0.091$$

Hence the probability of AIDS is only 9.1%!

Evidence from an Eye-Witness

A witness is 90% certain and there were 20 people at the crime scene . . .

$$\Pr(X|Y) = \frac{0.9 \cdot 0.05}{0.9 \cdot 0.05 + 0.1 \cdot 0.95} = 0.3213 = 32\%$$ now that's a worry . . .

Inference

Follow up on the AIDS test:

The doctor performs a, conditionally independent test which has the following properties:

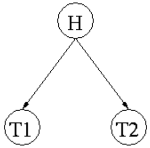
- The second test reports positive for 90% infections.
- The AIDS test reports positive for 5% healthy people.

$$\Pr(\text{T1}, \text{T2}|\text{Health}) = \Pr(\text{T1}|\text{Health}) \Pr(\text{T2}|\text{Health}).$$

A bit more algebra reveals $\frac{0.01 \cdot 0.05 \cdot 0.999}{0.01 \cdot 0.05 \cdot 0.999 + 1 \cdot 0.9 \cdot 0.001} = 0.357$.

Graphical Representation:

Through the unknown variable Health the outcomes of the two tests are coupled. We can view this via the following diagram:



Estimating Probabilities from Data

Rolling a dice:

Roll the dice many times and count how many times each side comes up. Then assign empirical probability estimates according to the frequency of occurrence.

Maximum Likelihood for Multinomial Distribution:

We match the empirical probabilities via

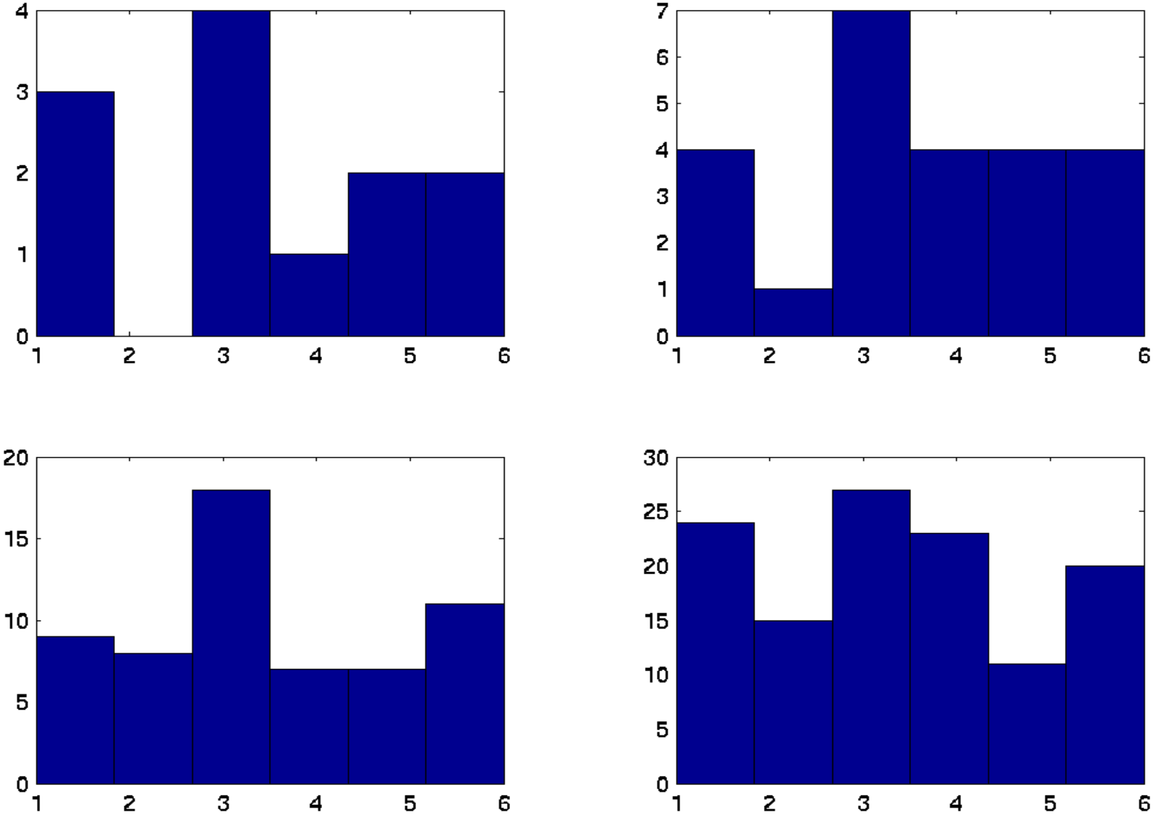
$$\Pr_{\text{emp}}(i) = \frac{\#\text{occurrences of } i}{\#\text{trials}}$$

Proof: we want to estimate the parameter vector $\pi \in \mathbb{R}^n$

$$\Pr(X|\pi) = \prod_{j=1}^{m} \Pr(X_j|\pi) = \prod_{i=1}^{n} \pi_i^{\#i}$$

Maximization subject to $0 \leq \pi_i$ and $\sum_i \pi_i = 1$ proves the claim.

Practical Example



Priors to the Rescue

Big Problem

Only sampling *many times* gets the parameters right.

Rule of Thumb

We need at least 10-20 times as many observations.

Priors

Often we know what we should expect. For instance assume a Dirichlet distribution over π, that is

$$\Pr(i|\pi) = \pi_i \text{ and } \Pr(\pi) \propto \prod_{i=1}^{n} \pi_i^{u_i - 1} \text{ where } u_i > 0.$$

Bayes rule yields $\Pr(\pi|X) \propto \prod_{i=1}^{n} \pi_i^{\#i + u_i - 1}$, which is maximized for $\pi_i = \frac{\#\text{occurrences of } i + u_i - 1}{\#\text{trials} + \sum_j (u_j - 1)}$. For $u_i = 2$ we obtain the Laplace Rule for estimation of frequencies.

An Outlook

Exponential Family

The multinomial distribution is a member of the exponential family where

$$\Pr(i|\pi) = \exp(\langle e_i, \log \pi \rangle - g(\pi))$$

Conjugate Prior

The Dirichlet prior is a conjugate prior for the multinomial family, i.e. $p(\pi)$ and $p(\pi|X)$ have the same form.

Translation: automatic way of finding “nice” priors.

Maximum a Posteriori Estimates

We chose π to maximize $p(\pi|X)$. This is also called the maximum-a-posteriori estimate.

Density Estimation

Data

Continuous valued random variables.

Naive Solution

Apply the bin-counting strategy to the continuum. That is, we use the empirical density $p_{\mathrm{emp}}(x) = \frac{1}{m}\sum_{i=1}^{m}\delta(x, x_i)$.

Problem

There are no bins.

Parzen Windows

Smooth out p_{emp} by convolving it with a kernel $k(x, x')$. Here $k(x, x')$ satisfies

$$\int_{\mathcal{X}} k(x, x')dx' = 1 \text{ for all } x \in \mathcal{X}.$$

Examples of Kernels

Gaussian Kernel

$$k(x, x') = \left(2\pi\sigma^2\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\|x - x'\|^2\right)$$

Laplacian Kernel

$$k(x, x') = \lambda^n 2^{-n} \exp\left(-\lambda\|x - x'\|_1\right)$$

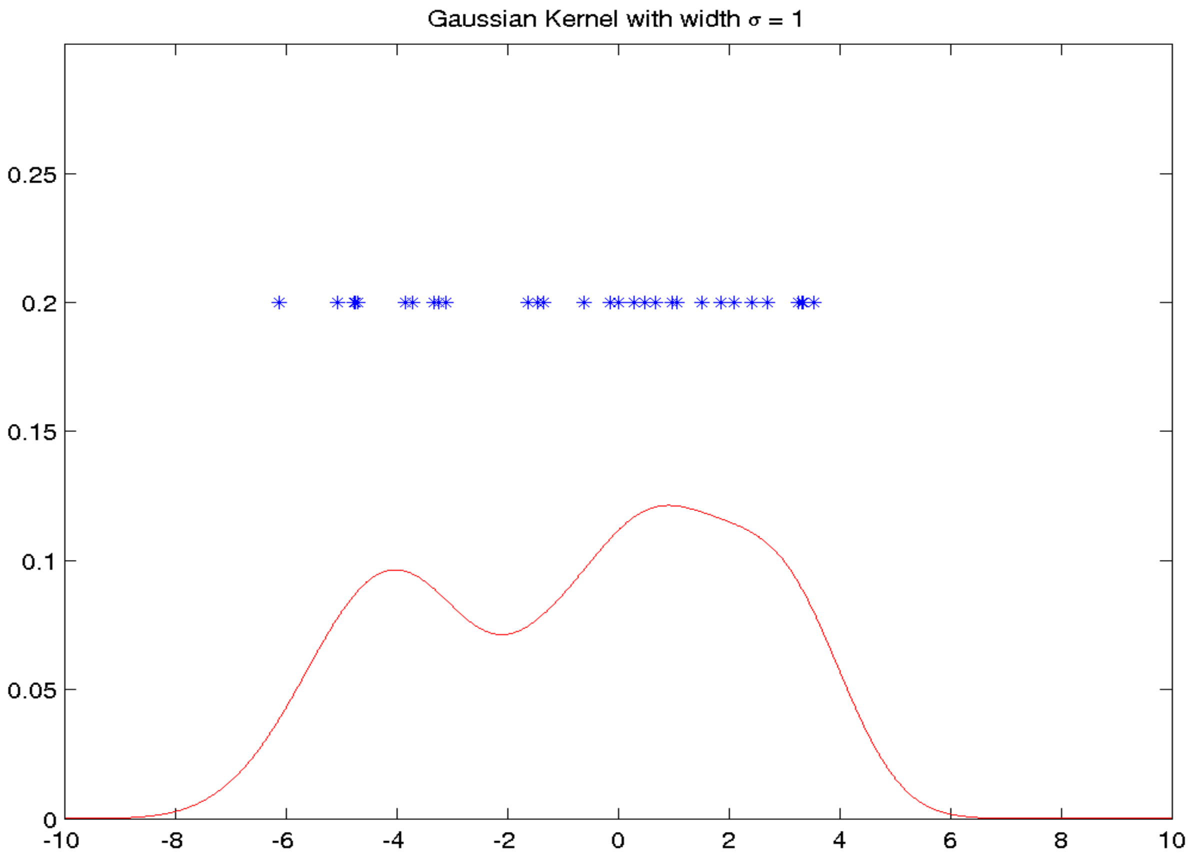
Indicator Kernel

$$k(x, x') = 1_{[-0.5,0.5]}(x - x')$$

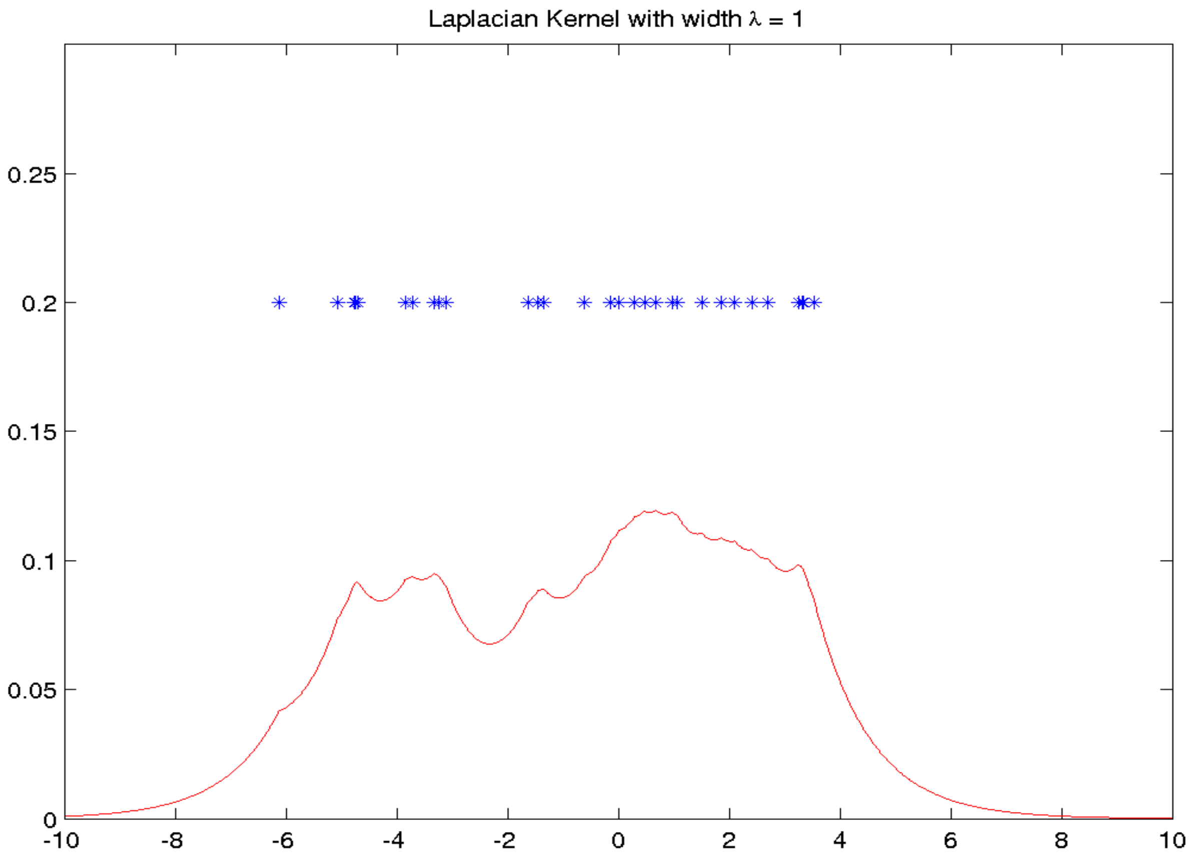
Important Issue

Width of the kernel is usually much more important than **type**.

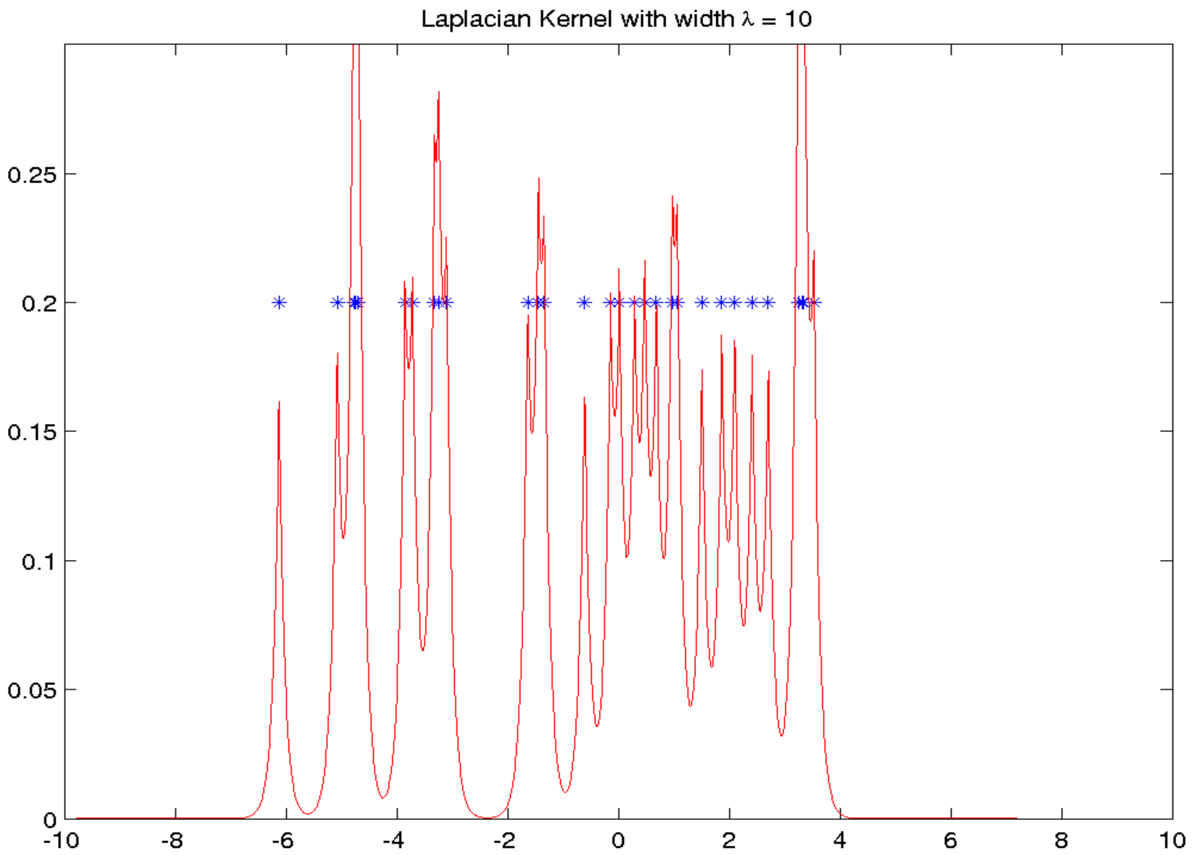
Gaussian Kernel



Laplacian Kernel



Laplacian Kernel



Selecting the Kernel Width

Goal

We need a method for adjusting the kernel width.

Problem

The likelihood keeps on increasing as we narrow the kernels.

Reason

The likelihood estimate we see is distorted (we are being overly optimistic through optimizing the parameters).

Possible Solution

Check the performance of the density estimate on an unseen part of the data. This can be done e.g. by

- Leave-one-out crossvalidation
- Ten-fold crossvalidation

Crossvalidation

Basic Idea

Compute $p(X'|\theta(X\setminus X'))$ for various subsets of X and average over the corresponding log-likelihoods.

Practical Implementation

Generate subsets $X_i \subset X$ and compute the log-likelihood estimate

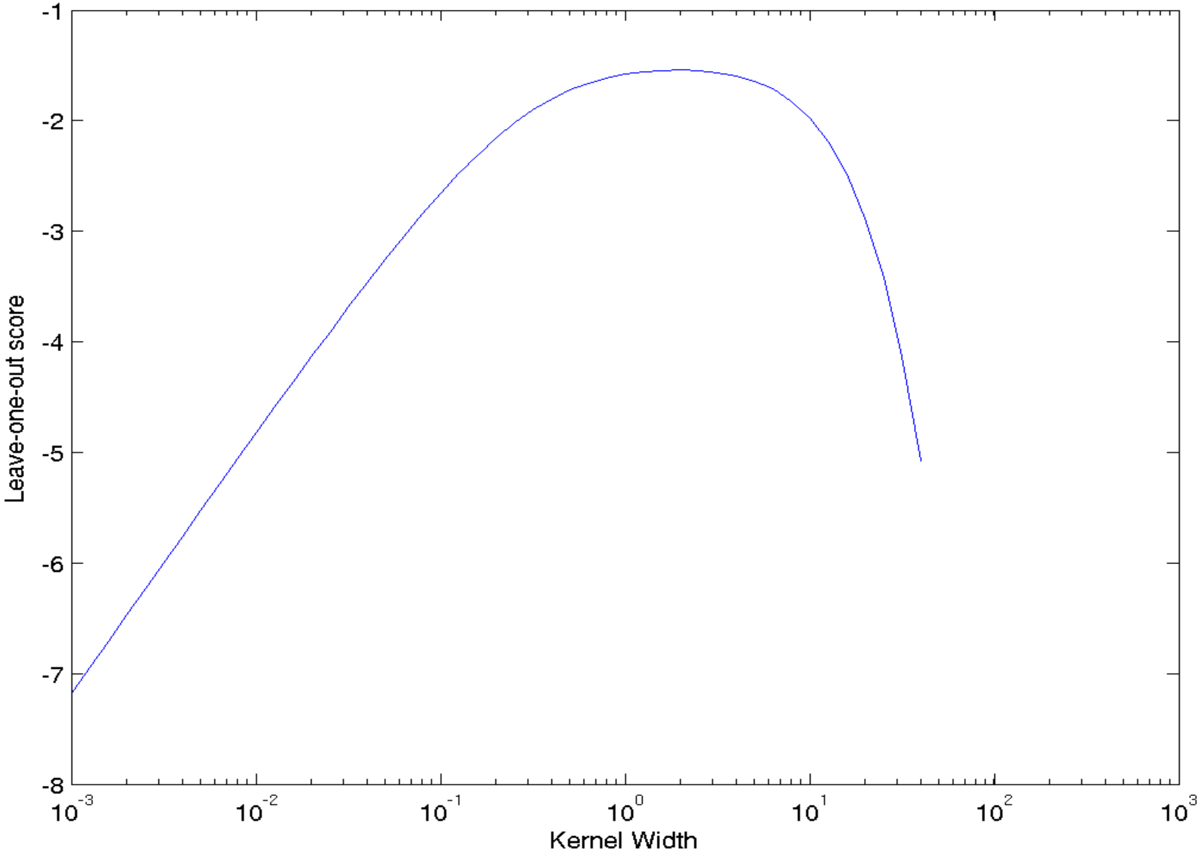
$$\sum_i \log p(X_i|\theta(X|\setminus X_i))$$

Pick the parameter which maximizes the above estimate.

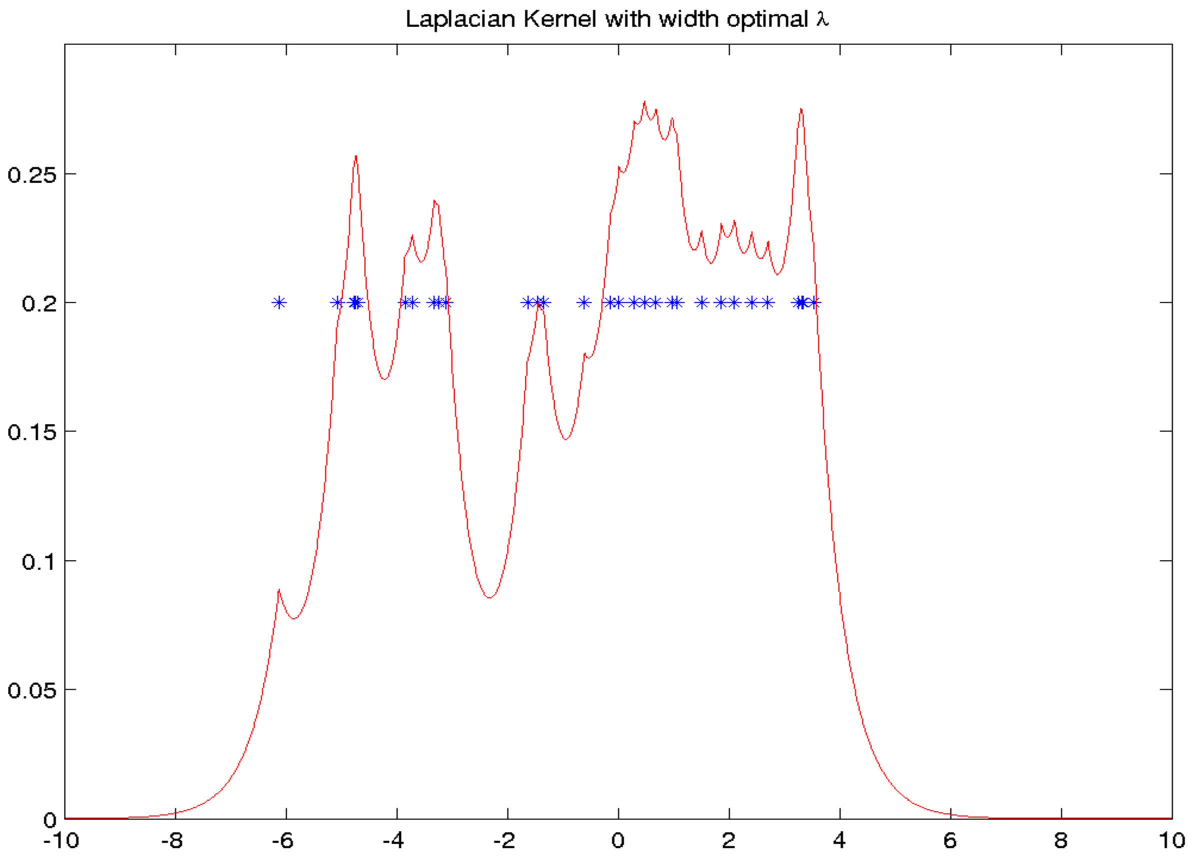
Special Case: Leave-one-out Crossvalidation

$$p_{X\setminus x_i}(x_i) = \frac{m}{m-1} p_X(x_i) - \frac{1}{m-1} k(x_i, x_i)$$

Cross Validation



Best Fit ($\lambda = 1.9$)



Application: Novelty Detection

Goal

Find the least likely observations x_i from a dataset X. Alternatively, identify low-density regions, given X.

Idea

Perform density estimate $p_X(x)$ and declare all x_i with $p_X(x_i) < p_0$ as novel.

Algorithm

Simply compute $f(x_i) = \sum_j k(x_i, x_j)$ for all i and sort according to their magnitude.

Applications

Network Intrusion Detection

Detect whether someone is trying to hack the network, downloading tons of MP3s, or doing anything else *unusual* on the network.

Jet Engine Failure Detection

You can't destroy jet engines just to see *how* they fail.

Database Cleaning

We want to find out whether someone stored bogus information in a database (typos, etc.), mislabelled digits, ugly digits, bad photographs in an electronic album.

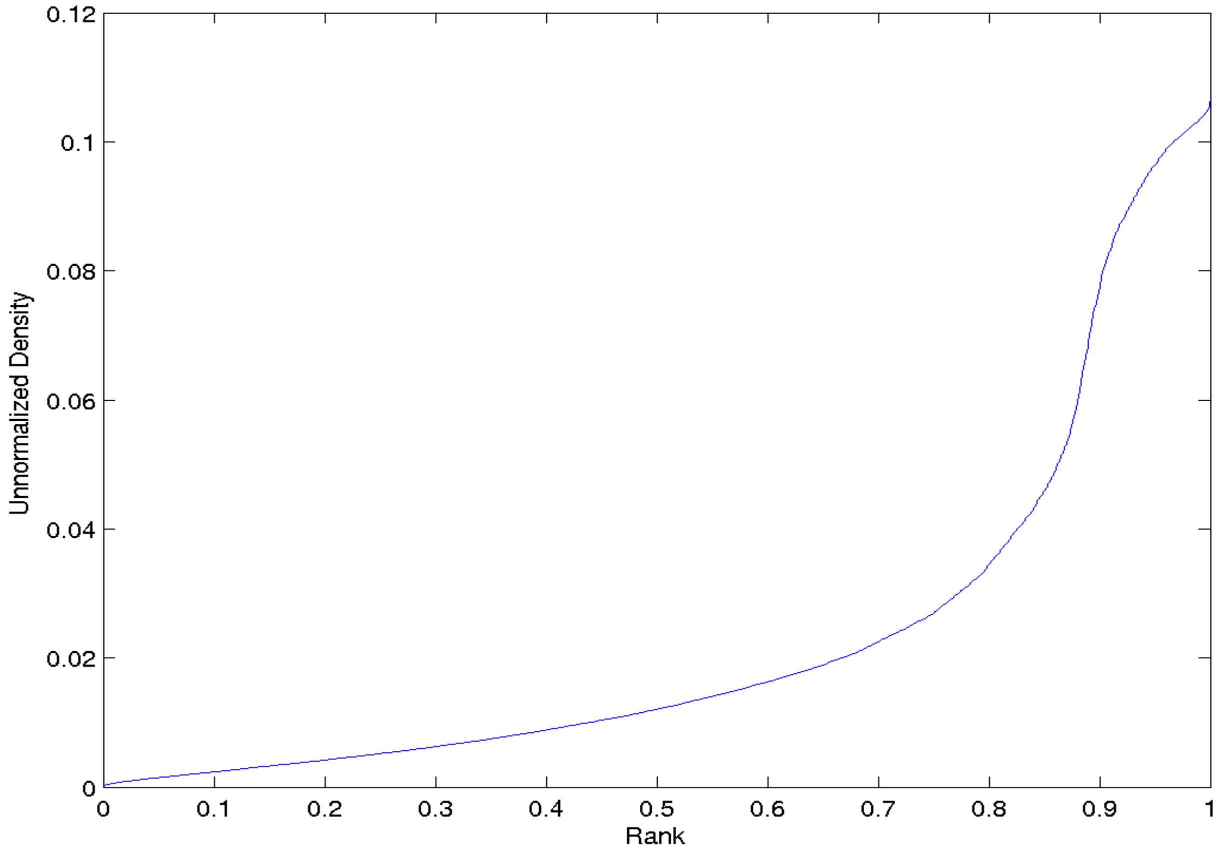
Fraud Detection

Credit Cards, Telephone Bills, Medical Records

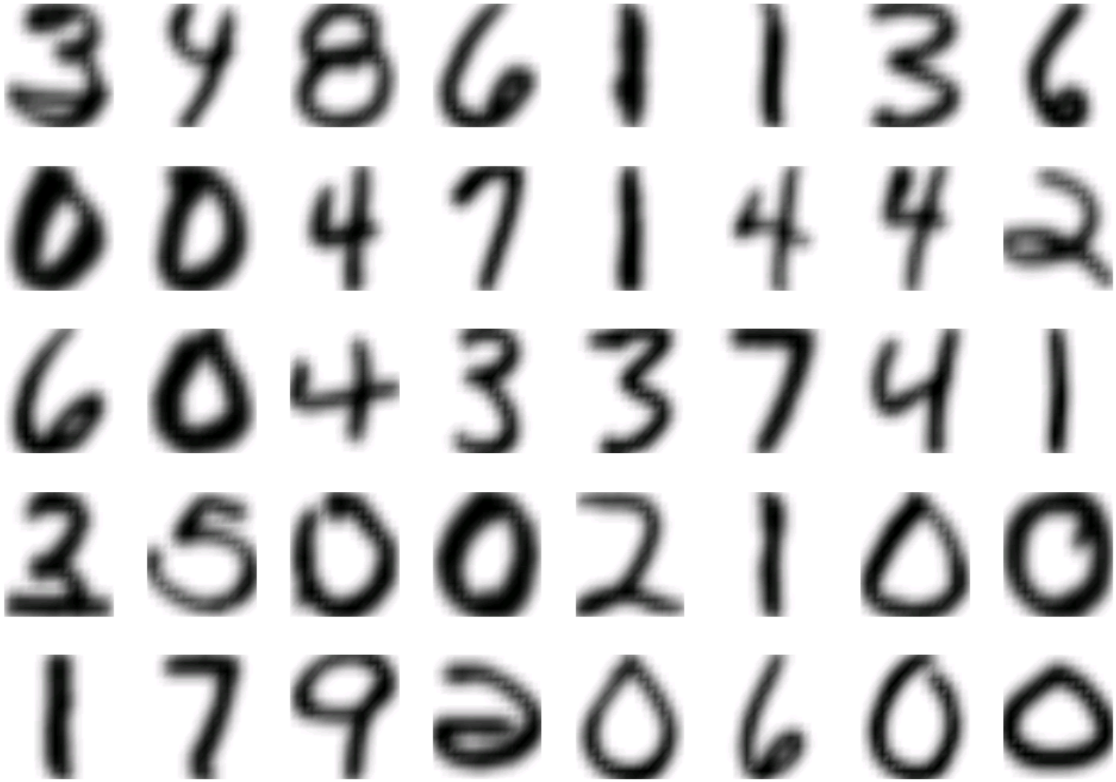
Self calibrating alarm devices

Car alarms (adjusts itself to where the car is parked), home alarm (furniture, temperature, windows, etc.)

Order Statistic of Densities



Typical Data



Outliers

Watson-Nadaraya Estimator

Goal

Given pairs of observations (x_i, y_i) with $y_i \in \{\pm 1\}$ find estimator for conditional probability $\Pr(y|x)$.

Idea

Use definition $p(x, y) = p(y|x)p(x)$ and estimate both $p(x)$ and $p(x, y)$ using Parzen windows. This yields

$$\Pr(y = 1|x) = \frac{\sum_{y_i=1} k(x_i, x)}{\sum_i k(x_i, x)}$$

Equivalent Formulation

Picking $y = 1$ or $y = -1$ depends on the sign of

$$\Pr(y = 1|x) - \Pr(y = -1|x) = \frac{\sum_i y_i k(x_i, x)}{\sum_i k(x_i, x)}$$

Extension to Regression

Use the above with $y_i \in \mathbb{R}$ for regression purposes.

NATIONAL ICT AUSTRALIA

Silverman's Automatic Adjustment

Problem

One 'width fits all' does not work well whenever we have regions of high and of low density.

Idea

Adjust width such that neighbors of a point are included in the kernel at a point. More specifically, adjust range h_i to yield

$$h_i = \frac{r}{k} \sum_{x_j \in \mathrm{NN}(x_i,k)} \|x_j - x_i\|$$

where $\mathrm{NN}(x_i, k)$ is the set of k nearest neighbors of x_i and r is typically chosen to be 0.5.

Result

State of the art density estimator, regression estimator and classifier.

Nearest Neighbor Classifier

Extension of Silverman's trick

Use the density estimator for classification and regression.

Simplification

Rather than computing a *weighted* combination of labels to estimate the label, use an *unweighted* combination over the nearest neighbors.

Result

k-nearest neighbor classifier. Often used as baseline to compare a new algorithm.

Nice Properties

Given enough data, k-nearest neighbors converges to the best estimator possible (it is consistent).