Pattern Analysis Joint Meeting
Christian Riess

Task 1: April 29

Agenda

- Task 1: Think-Pair-Share, about 25 minutes, consisting of:
  5 minutes introduction
  10 minutes in break-out rooms
  10 minutes plenary session

  Remarks:
  - Let us try out zoom's whiteboard for collecting the replies (→ share screen → white board).
  - The "host" will have to have a relatively stable internet connection.
  - The "writer" will have to have a relatively stable internet connection.
  - Virtually "raise your hand" to contribute, unmute when host calls you. Don't be afraid of wrong answers, we can fix errors later.
  - Note: every participant can save the content of the white board as an image, if you like we can upload a copy to studon.
- Task 2: Program adaptation in a Think-Pair-Share format, about 25 minutes, consisting of:
  5 minutes introduction
  10 minutes in break-out rooms
  10 minutes plenary session

  Remarks:
  - We will again need a host with a relatively stable internet connection.
  - Raise your hand to say something or to show something. Wait for host to call you, then unmute.
  - Let us try out screen sharing of participants to show around solutions. Screen sharers will have to have a relatively stable internet connection.
- Task 3: Think-Pair-Share, about 35 minutes or as long as time permits
  5 minutes introduction
  15 minutes in break-out rooms
  15 minutes plenary session

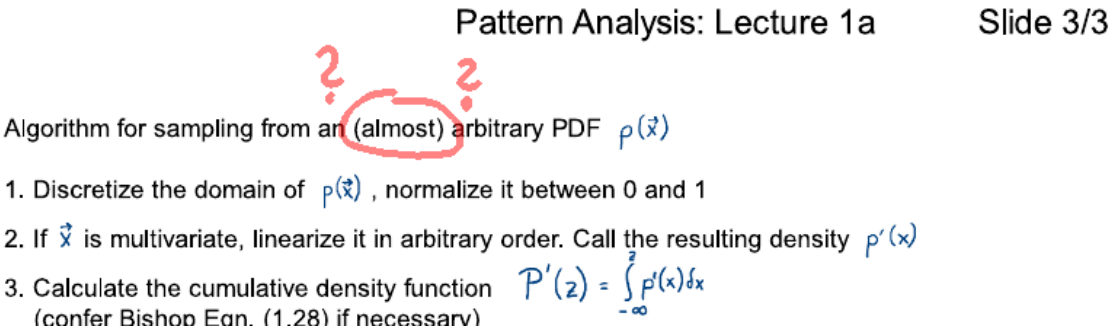
  Remarks:
  - Same game as task 1. If the white board worked for task 1 let us try this again.

Pattern Analysis Joint Meeting
Christian Riess

Task 1: April 29

Task 1

Let us discuss a tiny word in the lecture notes, section 00a. There is this weird remark when presenting the algorithm for sampling from a PDF:



Pattern Analysis: Lecture 1a Slide 3/3

Algorithm for sampling from an (almost) arbitrary PDF $p(\vec{x})$

1. Discretize the domain of $p(\vec{x})$, normalize it between 0 and 1
2. If $\vec{x}$ is multivariate, linearize it in arbitrary order. Call the resulting density $p'(x)$
3. Calculate the cumulative density function $P'(z) = \int_{-\infty}^{z} p'(x)\,dx$ (confer Bishop Eqn. (1.28) if necessary)

What does "almost" refer to? Is there a theoretical issue? Or are there possible practical issues?

Think of a potential failure case of this method.

Pattern Analysis Joint Meeting
Christian Riess

Task 1: April 29

Task 2

Think of a concrete function where your implementation of the sampling does not give the expected result.

If you have time, think for a minute about the question what a particularly nice function might be to test this.

→ Just for fun, hack either of these functions into your code, and create an output image!

Pattern Analysis Joint Meeting
Christian Riess

Task 2: May 6

Task 1: Kernel Forensics

Assume that we examine the PDF that is produced by one of our density estimators.

What can we find out about the used density estimation technique?

Specifically:

(a) Can we distinguish different kernel functions for density estimation?
If yes: does it always work? What are necessary and/or sufficient conditions for success?

(b) Can we make a guess on the kernel window size?
If yes: does it always work? What are necessary and/or sufficient conditions for success?

(c) Please also include in your discussion the k-NN density estimator!

Pattern Analysis Joint Meeting
Christian Riess

Task 2: May 6

Task 2: Computational Cost and Storage Cost

Assume that we have a d-dimensional feature space (d is potentially large), and that — without loss of generality — all samples are located in a d-dimensional hypercube with edge length 1 (i.e., the volume is also 1). We use a kernel with compact support[1].

(a) Discretize each sample dimension into B bins. How much storage does the PDF need?

(b) To improve the resolution: how much more storage do you need to double the number of bins?

(c) What is the computational cost for calculating this PDF?

(d) Let us assume that we do not want to discretize the space. Instead, we would like to return an exact (non-discretized) density estimate at position x_0.
How does this affect our computational requirements?
What might be possible ways to reduce the computational cost?

As always: feel free to contribute to your breakout room knowledge and skills from your other fields of study if it helps to answer these questions!

[1] *compact support* means that only finitely many entries are non-zero, such that the notion of "window size" applies

Pattern Analysis Joint Meeting
Christian Riess

Worksheet 3: May 13

Task 1: All Interchangeable: Density Estimation, Classification, and Regression?

The initial statement in lecture 01e is that density estimation naturally links to classification and regression. In other words, density estimation can be converted to regression or to classification.

Let us discuss this a little bit more, to make sure that we all feel very smart about this.

(a) Density estimation versus regression:

  (a) Please clarify with your room members both directions: how to emulate a PDF via regression, and how to emulate a regression estimate via a probability density estimator.

  For the first part (emulation of a PDF): how can we sample from that PDF?

  (b)

(b) Density estimation versus classification:

  (a) Please clarify with your room members both directions: how to emulate a PDF via a classifier, and how to emulate a classifier with a PDF.

  For the first part (emulation of a PDF): how can we sample from that PDF?

(c) In light of the previous discussion:

  What advantages might be gained if you have a dedicated regressor, classifier, or density estimator instead of "jury-rigging" one from the respective other task?

Pattern Analysis Joint Meeting
Christian Riess

Worksheet 3: May 13

Task 2: Higher-order Correction Terms

You successfully applied this week a linear correction factor to the kernel estimate to fix a first order bias of the estimate, which may occur, e.g., at boundaries of the sample domain.

The following chapter, HTF Sec. 6.1.2, states that we are not limited to linear corrections: any degree of a polynomial could be applied by extending the linear fit (Eqn. (6.7) on page 195)

$$\min_{\alpha(x_0),\beta(x_0)} \sum_{i=1}^{N} K_\lambda(x_0, x_i)(y_i - \alpha(x_0) - \beta(x_0)x_i)^2 \tag{1}$$

to a polynomial fit (Eqn. (6.11) on page 197), namely

$$\min_{\alpha(x_0),\beta(x_0)} \sum_{i=1}^{N} K_\lambda(x_0, x_i) \left(y_i - \alpha(x_0) - \sum_{j=1}^{d} \beta_j(x_0) x_i^j \right)^2 . \tag{2}$$

(a) How do you need to change your code to accomodate for this extension?

(b) Difficult question 1: HTF state that this reduces the bias to terms of order d+1, but at the price of an increased variance. What does that mean? Can you rephrase this for your room mates?

(c) Difficult question 2: can you quickly hack the polynomial fit into your code, such that we can look at the result?

There is some wisdom that can be harvested from question (b). Thus, don't worry if you can not find a solution, we will make sure that we have a result on this in the plenary session.

Pattern Analysis Joint Meeting
Christian Riess

Worksheet 4: May 20

Task 1: Random Forests: Tree Depth and Forest Size

Consider a Random Forest trained on the raccoon image (last exercise).

(a) One parameter that is oftentimes varied is the maximum tree depth. How many leaves has a tree with depths 10, 20, or 50?

(b) Estimate how large the partition of a raccoon image is for tree depths of 10, 20, 50?

(c) Give a rough estimate how much computer memory you need in an efficient low-level representation for such a tree of depth 10, 20, or 50.

(d) Will your python code crash if you set the maximum depth to 50?

(e) Generally speaking, more trees help to reduce overfitting. Is this also true for an ensemble of trees of maximum depth 50?

Pattern Analysis Joint Meeting
Christian Riess

Worksheet 5: May 27

Task 1: Clustering Forensics

Your friend gives you a picture of a 2-D feature space. In that feature space, sample points are colored to indicate their cluster membership. Unfortunately, you don't know which method was used to cluster the data.

Is there a way how you could distinguish a mean shift clustering from a k-Means clustering? If yes, clarify the circumstances and/or conditions under which they are distinguishable. If no, clarify the reasons why not.

Follow-up question: is there a way to distinguish a k-Means clustering from a GMM clustering, if the GMM clustering always assigns that clustering label with the highest probability? As above, explain your "yes" or "no" answer.

Pattern Analysis Joint Meeting
Christian Riess

Worksheet 5: May 27

Task 1: Closing Some Gaps in the Gap Statistics

Let us discuss a few details kMeans, the gap statistics, and mean shift model selection.

(a) There is a question in the results thread of the forum on the optimality of kMeans clusterings. kMeans clusterings are only locally optimal, i.e., the solution depends on the initialization. This property may be annoying for calculating the gap statistics. One of the supervisors recommends to average a couple of runs — which is one possibility. In fact, any other statistics could also be chosen, such as the minimum of a couple of runs, or some quantile, ...

Make some general considerations: when might it preferable to take the average, when a minimum?

(b) What purpose(s) serves the standard deviation s'_{K+1} in the equation of the gap statistics? For your convenience, the equation is

$$K^* = \underset{K}{\operatorname{argmin}}\{K | G(K) \geq G(K+1) - s'_{K+1}\} \ , \tag{1}$$

where $G(K) = \log(W_K^{\text{unif}}) - \log(W_K^{\text{data}})$

(c) A cross-link to mean shift: Comaniciu and Meer present a wealth of general approaches for optimizing the kernel bandwidth. These methods should also apply to mean shift. Option 3 proposes to use some metrics on the cluster quality like the within-cluster distance. However, I am sceptical whetherthat this is necessarily a suitable choice for mean shift.

What could be a concrete concern when using the within-cluster distance for optimizing the mean shift bandwidth?

Pattern Analysis Joint Meeting
Christian Riess

Worksheet 7: June 17

Task 1: Clarifications on Manifold Learning

Let us do just three straightforward questions today:

(a) Why are PCA and MDS called "linear" methods?

(b) Why do we consider the largest eigenvalues for PCA, MDS, ISOMAP, and the smallest eigenvalues for Laplacian Eigenmaps?

(c) ISOMAP is known to be more sensitive to noise than Laplacian Eigenmaps. What might be the reason for this?

Pattern Analysis Joint Meeting
Christian Riess

Worksheet 8: June 24

Task 1: Clarifications on Manifold Learning

Let us do three questions today:

(a) Please clarify: which are the main factors that lead to a smooth density in the Density Forest?

(b) A question from the forum: why do we draw samples from a random forest by first picking a forest, then randomly walk down the nodes, then draw a sample from that Gaussian? Wouldn't it be simpler to directly sample from the extrapolated CDF?

(c) When might we have a benefit from using a Density Forest or a Manifold Forest over a kernel-based density estimator or the vanilla Laplacian Eigenmaps? When would we not have that benefit?