Discriminative Learning and Big Data

Lecture 2: Multiple Classes, Retrieval, and Approximate Nearest Neighbours

Andrew Zisserman and Relja Arandjelović

Visual Geometry Group
University of Oxford
http://www.robots.ox.ac.uk/~vgg
Lecture outline

• Beyond Binary Classification
  • Multi-class and Multi-label
  • Using binary classifiers

• Big Data
  • retrieval and ranking
  • precision-recall curves

• Part II
  • Approximate Nearest Neighbours (ANN)
  • Introduction to practical
Multi-class vs Multi-label Classification
Multi-class vs Multi-label

Have K classes: \{C_1, C_2, \ldots, C_K\}

1. Multi-class classification (exclusive)
   - each \(x\) assigned to one (and only one) label
   - e.g. recognize digits for zip codes (1 of 10 choices at each position)

2. Multi-label classification (non-exclusive)
   - each \(x\) assigned to a set of labels
   - e.g. annotate an image with the object categories it contains
     -> \{car, horse, person\}
Multi-class vs Multi-label

e.g. K = 5 classes, and one-hot encoding

1. Multi-class classification (exclusive)
   • each x assigned to one (and only one) label

2. Multi-label classification (non-exclusive)
   • each x assigned to a set of labels

<table>
<thead>
<tr>
<th>data</th>
<th>labels j</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 1 0 0 0</td>
<td>0 1 0 0 0</td>
</tr>
<tr>
<td>0 0 0 1 0</td>
<td>0 0 0 1 0</td>
</tr>
<tr>
<td>0 0 1 0 0</td>
<td>0 0 1 0 0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>data</th>
<th>labels j</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 1 0 0 1</td>
<td>0 1 0 0 1</td>
</tr>
<tr>
<td>1 1 0 1 0</td>
<td>1 1 0 1 0</td>
</tr>
<tr>
<td>0 0 1 0 0</td>
<td>0 0 1 0 0</td>
</tr>
</tbody>
</table>
Multi-class Classification
Multi-Class Classification – what we would like

Assign input vector $\mathbf{x}$ to one of $K$ classes $\mathcal{C}_k$

Goal: a decision rule that divides input space into $K$ decision regions separated by decision boundaries
Reminder: K Nearest Neighbour (K-NN) Classifier

Algorithm

- For each test point, $x$, to be classified, find the $K$ nearest samples in the training data
- Classify the point, $x$, according to the majority vote of their class labels

\[ e.g. \ K = 3 \]

- Naturally applicable to multi-class case
- NB beware that $K$ is number of NN here, not number of classes
Build from binary classifiers

- Learn: K two-class 1-vs-the-rest classifiers $f_k(x)$
Build from binary classifiers continued

- **Learn**: $K$ two-class 1-vs-the-rest classifiers $f_k(x)$
- **Classification**: choose class with most positive score

![Diagram showing classification with binary classifiers for classes C1, C2, and C3.](image)
Build from binary classifiers continued

- **Learn**: $K$ two-class 1 vs the rest classifiers $f_k(x)$
- **Classification**: choose class with most positive score

$$\max_k f_k(x)$$
Application: hand written digit recognition

- **Feature vectors**: each image is 28 x 28 pixels. Rearrange as a 784-vector $\mathbf{x}$

- **Training**: learn $k=10$ two-class 1-vs-the-rest SVM classifiers $f_k(x)$

- **Classification**: choose class with most positive score

$$f(x) = \max_k f_k(x)$$
Example

hand drawn

classification

STPRTool  Franc & Hlavac
Why not learn a multi-class SVM directly?

For example for three classes

- Learn \( w = (w_1, w_2, w_3)^\top \) using the cost function

\[
\min_w \|w\|^2 \quad \text{subject to}
\]

\[
w_1^\top x_i \geq w_2^\top x_i \quad \& \quad w_1^\top x_i \geq w_3^\top x_i \quad \text{for } i \in \text{class 1}
\]

\[
w_2^\top x_i \geq w_3^\top x_i \quad \& \quad w_2^\top x_i \geq w_1^\top x_i \quad \text{for } i \in \text{class 2}
\]

\[
w_3^\top x_i \geq w_1^\top x_i \quad \& \quad w_3^\top x_i \geq w_2^\top x_i \quad \text{for } i \in \text{class 3}
\]

- This is a quadratic optimization problem subject to linear constraints and there is a unique minimum

- Note a margin can also be included in the constraints.

In practice there is a little or no improvement over the binary case.
Reminder: Logistic Regression (LR) binary classifier

- Binary labels $y_i \in \{0, 1\}$

- LR Classifier:
  \[
  p(y = 1|x; w) = \sigma(f(x)) = \frac{1}{1 + e^{-f(x)}}
  \]
  where $f(x) = w^\top x + b$.

- Learning Loss:
  \[
  L(w) = -\sum_{i=1}^{N} y_i \log \sigma(f(x_i)) + (1 - y_i) \log (1 - \sigma(f(x_i)))
  \]
  Note,
  1. This is the cross-entropy between the ground truth and predicted distributions
  2. It is a convex function
Logistic Regression to Soft-Max

LR naturally generalizes to multiple classes using a Soft-Max

\[ P(C_k|x) = \frac{\exp(f_k(x))}{\sum_j^K \exp(f_j(x))} \]

where \( f_k(x) = w_k^\top x + b_k \)

Note:

- \( 0 \leq P(C_k|x) \leq 1 \) and \( \sum_k^K P(C_k|x) = 1 \)

- Behaves as a soft-max since if \( f_k(x) \gg f_j(x) \) for all \( j \neq k \)
  then \( P(C_k|x) \approx 1 \) and \( P(C_j|x) \approx 0 \)

Helpful post by Chris Yeh clarifying relation:
Soft-Max classifier/Cross-entropy loss

- **K** classes, \( \{C_1, ..., C_k, ... C_K\} \)

- **Classifier:** Soft-max

\[
P(C_k | x) = \frac{\exp(f_k(x))}{\sum^K_j \exp(f_j(x))}
\]

where \( f_k(x) = w_k^\top x + b_k \)

- **Loss function:** is the negative logarithm of the estimated class probability of the true class \( k \):

\[
- \log \left( \frac{\exp(f_k(x))}{\sum^K_j \exp(f_j(x))} \right) = -f_k(x) + \log \sum^K_j \exp(f_j(x))
\]

Note, this is the **cross-entropy** between the distribution of the estimated class probabilities and the true distribution, i.e. a probability of one for the correct class and zero for the others.
Multi-label classification

- Again, can use binary classifiers, such as one-vs-rest SVM or LR, one for each class

- e.g. with a LR for each data point $x$, the loss function is

$$- \sum_{k}^{K} y_k \log \sigma(f_k(x)) + (1 - y_k) \log(1 - \sigma(f_k(x)))$$

where $y_k \in \{0, 1\}$ is the ground truth for class $k$.

- Evaluation measures:
  - Hamming distance: the fraction of the wrong labels to the total number of labels
  - Intersection over Union (IoU) (Jaccard Index): the number of correctly predicted labels divided by the union of predicted and true labels
Cost functions are defined for many other tasks, e.g. for embedding:

- Contrastive loss: small if two samples from the same class, large otherwise. Used in verification.

- Triplet loss: distance of an anchor to positive less than distance to negative by a margin. Used in ranking.
Big Data means at least …

- Image search, e.g. Facebook
  - 1 billion images = 1 billion descriptors (e.g. 128D based on CNNs)

- Video search: thousands of hours of video
  - Billions of audio and video descriptors

- Music search: e.g. Shazam

- The order of magnitude considered in this lecture is millions to billions
Example: Large scale image/video retrieval

**Objective:** retrieve images from an image or video dataset that contain a particular object category

**Offline:**
Represent every image (or keyframe) in dataset by a feature vector

**Online:**
**Query vector:** Train an SVM classifier using positive images that contain the category and negative images that don’t

- Score each image in the dataset by the classifier, and rank by score
- Basic operation: nearest neighbour matching/retrieval
Image retrieval

**Training**
- **Training Images**
  - Positive (Airplane)
  - Negative (Background)
- **Image encoding**
- **Features**
- **Learning**
  - **Classifier Model**
    - Linear SVM

**Retrieval**
- **Dataset Images**
- **Image encoding**
- **Features**
- **Scoring**
  - Scores:
    - 0.91
    - 0.12
    - 0.65
    - 0.89
- **Ranking**
  - Ranked List:
    - 0.91
    - 0.89
    - 0.65
    - 0.12
Video dataset: BBC TV

- 17K broadcasts from BBC 1, 2, 3, 4 & BBC News 24
- Programmes from 2007 to 2012 in prime time slot (7-12pm)
- 10K hours of video represented by 1 frame per second
- 36M seconds of data, 5M keyframes
- Frames are 480 x 270 pixels
Two Aspects of Information Retrieval

1. How to evaluate the quality of the retrieved results:
   - Precision-Recall (PR) curves

2. How to efficiently retrieve from large scale datasets
   - Approximate nearest neighbour search
Precision – Recall curve

- **Precision**: fraction of the retrieved set that are positives
- **Recall**: fraction of all positives in retrieved set

Classifier score decreasing
Precision-Recall curve

- **Precision**: fraction of the retrieved set that are positives
- **Recall**: fraction of all positives in retrieved set

Number of true positives = 8
A pot-pourri of PR curves
Average Precision (AP)

- Area under Precision-Recall curve
- Single score to assess performance

AP = 0.23
AP = 0.44

A good score requires both high recall and high precision
Other Information Retrieval performance measures

• **Precision at k (Pr@k)**
  • The number of positives in the top k

<table>
<thead>
<tr>
<th>Rank</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
</table>

• **Discounted Cumulative Gain (DCG)**
  • Allows for relevance (rel) weighting
  • DCG at rank p is defined as:

\[
DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{\log_2(i + 1)}
\]

• Normalized DGC (nDCG):

\[
nDCG_p = \frac{DCG_p}{IDCG_p}
\]

• where IDCG is the DCG of a perfect ranking and lies between [0,1]
Nearest Neighbours

Given a query vector $q$ find the closest vector $x$ in the dataset

$$\text{NN}(q) = \arg \min_x \|x - q\|^2$$

- $d$ dimensional vectors
- $n$ size of dataset

NN and linear classifiers:

Given a classifier $w$ score each vector $x$ in the dataset as $w \cdot x$ and sort

$$\|w - x\|^2 = \|w\|^2 + \|x\|^2 - 2w \cdot x = 2 - 2w \cdot x$$

with $x$ and $w$ having a unit norm.
Nearest Neighbours

Given a query vector $q$ find the closest vector $x$ in the dataset

$$\text{NN}(q) = \arg \min_x ||x - q||^2$$

- $d$ dimensional vectors
- $n$ size of dataset

Problems:
- Costly operation of exact exhaustive (linear) search: $O(n*d)$
- High-dimensional vectors: for exact search the best approach is the naïve exhaustive comparison

Related task: k-NN
- find $K$ nearest neighbours
- used in retrieval for ranked lists and for classification
The cost of (efficient) exact matching

But what about the actual timings? With an efficient implementation!

Finding the 10-NN of 1000 distinct queries in 1 million vectors
  • Assuming 128-D Euclidean descriptors
  • i.e., 1 billion distances, computed on a 8-core machine

How much time?
The cost of (efficient) exact matching

But what about the actual timings? With an efficient implementation!

Finding the 10-NN of 1000 distinct queries in 1 million vectors
  - Assuming 128-D Euclidean descriptors
  - i.e., 1 billion distances, computed on a 8-core machine

5.5 seconds
Need for approximate nearest neighbours

To improve the scalability:

• **Approximate nearest neighbor (ANN) search**

Three (contradictory) performance criteria for ANN schemes

• search quality (retrieved vectors are actual nearest neighbors)
• speed
• memory footprint

Representation must fit into memory (disk is too slow)
Approximate Nearest Neighbours

Relja Arandjelović

DeepMind
www.relja.info
NN Applications: What is this?

Search for the query descriptor in a large database
Copy the meta-data of the nearest neighbour

HOW DOES SHAZAM WORK?

This is a question we get often asked. Here is a quick summary of the three main steps involved from the moment you Shazam until the magic happens.

Let’s say you are in a shop and you like the music you’re hearing. Start the app and tap the Shazam button.

A digital fingerprint of the audio is created and, within seconds, matched against Shazam's database of millions of tracks and TV shows.

You are then given the name of the track and the artist and information such as lyrics, video, artist biography, concert tickets and recommended tracks. We also let you buy or listen to the song using one of our partners’ services.

From www.shazam.com
NN Applications: Automatic panorama creation

Match local patches, estimate the geometric transformation

From Brown & Lowe "Automatic Panoramic Image Stitching using Invariant Features"
NN Applications: Reinforcement Learning

Act in the environment by consulting past memories
Search memories using the current state

From Min & Kim “Learning to Play Visual Doom using Model-Free Episodic Control”
NN Applications: Data Visualization (t-SNE)

Compute nearest neighbours for all points
Flatten the high dimensional data preserving the neighbourhood structure

Taken from https://theconceptcenter.com/simple-research-study-similarity-between-two-words/
NN Applications: Other

- k-NN classification and regression
- Clustering
- Hard negative mining
- Semi-supervised learning

...
Efficient implementations

Cost $O(nd)$ but many optimization tricks
- Scalar product instead of distance if possible
- Special CPU instructions (SIMD in SSE)
- Pipelining / loop expansion
- Multiple queries: clever use of cache

For best performance use standard libraries
- Lower-level: BLAS
- Higher-level: Vectorized MATLAB, numpy, ..
- Good free libraries:
  - yael_nn (http://yael.gforge.inria.fr/)
  - FAISS (https://github.com/facebookresearch/faiss)
The cost of (efficient) exact matching

But what about the actual timings? With an efficient implementation!

Finding the 10-NN of 1000 distinct queries in 1 million vectors
• Assuming 128-D Euclidean descriptors
• i.e., 1 billion distances, computed on a 8-core machine

How much time?
The cost of (efficient) exact matching

But what about the actual timings? With an efficient implementation!

Finding the 10-NN of 1000 distinct queries in 1 million vectors
- Assuming 128-D Euclidean descriptors
- i.e., 1 billion distances, computed on a 8-core machine

5.5 seconds
How to speed up?

Dimensionality reduction?
  • Reasonable first step, but typically insufficient

Use GPU?
  • Adds lots of complexity
  • Insufficient memory
  • Overhead for memory copying between CPU & GPU

Buy more machines?
  • Costs money to buy, maintain, ..
  • Adds lots of complexity
  • For real-time systems: communication overhead
  • Still often insufficient, e.g. if all pairwise distances are needed (e.g. building a neighbourhood graph, clustering, ..)
Finding *approximate* nearest neighbour vectors

- Approximate method is not guaranteed to find the nearest neighbour.

- Can be much faster, but at the cost of missing some nearest matches

![Diagram showing approximate nearest neighbour space in 128D descriptor space](image)

- Found (near) match
- Query
- True NN match
- 128D descriptor space
Approximate Nearest Neighbours (ANN)

Is finding only approximate nearest neighbours acceptable?

• Often there is no choice
  • Use ANN or not = have Google or not

• Often it is good enough
  • What is this?

? Big Ben
Approximate Nearest Neighbours (ANN)

Is finding only approximate nearest neighbours acceptable?

- Often there is no choice
  - Use ANN or not = have Google or not
- Often it is good enough
  - What is this?

Big Ben
Approximate Nearest Neighbours (ANN)

Is finding only approximate nearest neighbours acceptable?

- Often there is no choice
  - Use ANN or not = have Google or not

- Often it is good enough
  - What is this?
  - Geometric matching: Only need sufficient number of matches
  - Data visualization: Rough nearest neighbours are fine
Approximate Nearest Neighbours (ANN)

Three (contradictory) performance criteria for ANN schemes

• Search quality
  • Retrieved vectors are actual nearest neighbors

• Speed

• Memory footprint
  • Representation must fit into memory, disk is too slow
Approximate Nearest Neighbours (ANN): Methods

Broad division

1. Approximate the vectors => do faster distance computations
   • Brute-force search, visit all points: $O(n)$

2. Approximate the search => do fewer distance computations
   • Do not visit all points

Can (and should) mix and match: approximate both, rerank (1.) with original vectors, etc
Approximate Nearest Neighbours (ANN): Methods

Broad division

1. **Approximate the vectors** => do faster distance computations
   - Brute-force search, visit all points: $O(n)$

2. **Approximate the search** => do fewer distance computations
   - Do not visit all points
ANN via approximating vectors

- Objectives
  - Approximate vectors as well as possible
  - Make the distance computation fast
  - Effectively: quantization, compression

- Do we need all 32 bits in a float?
  - Sometimes yes:
  - Sometimes no:
  - The limit is entropy (amount of information)
  - Betting on exploiting underlying structure
    - Otherwise there is no hope
    - zip, jpeg, mpeg, .. do not always compress!
ANN via approximating vectors: Methods

- Hashing
  - LSH

- Vector Quantization
  - Product Quantization
Hashing (for approximating vectors)

- **Key idea**
  - Convert the input real-valued vector into a binary vector
  - Compare binary vectors using Hamming distance
    
    \[
    x: 00100101 \\
    y: 10100011
    \]
    
    \[10000110 \Rightarrow \text{Hamming distance} = \text{sum(1's)} = 3\]

- **Pros**
  - Memory savings: \(d \times 32 \text{ bits} \rightarrow b \text{ bits}\)
  - Extremely fast distance computation: \(\text{popcnt}(\text{xor}(x,y))\)

- **Cons**
  - Can be hard to design a good hashing function
Locality Sensitive Hashing (LSH)

- Choose a random projection
- Project points, pick a random threshold
- Points close in the original space remain close under the projection
- Unfortunately, converse not true
Locality Sensitive Hashing (LSH)

- Choose a random projection
- Project points, pick a random threshold
- Points close in the original space remain close under the projection
- Unfortunately, converse not true
- Solution: use multiple quantized projections which define a high-dimensional “grid”
Locality Sensitive Hashing (LSH): Discussion

- Pros
  - Very simple to implement
  - Very fast
  - Good for embedded systems
Locality Sensitive Hashing (LSH): Discussion

• Pros
  • Very simple to implement
  • Very fast
  • Good for embedded systems

“Choose a random projection” ??
“Betting on exploiting underlying structure”

• Cons
  • Data-agnostic = blind
  • Need a large number of projections (=bits=a lot of memory) for decent performance compared to alternatives

• The way to go: data-dependent hashing
  • Iterative Quantization [Gong & Lazebnik 2011], Spectral Hashing [Weiss et al. 2009], …
ANN via approximating vectors: Methods

- Hashing
  - LSH

- Vector Quantization
  - Product Quantization
Vector Quantization

Hard to design the conversion real vector -> binary vector, such that distances in binarized space are meaningful.

Idea

- Scrap using Hamming distance to compare codes
- Codebook (vocabulary) of $k$ vectors (codewords)
- Code for vector = ID of the nearest codeword ($\log_2 k$-bits)
- Learn codebook to optimize for quantization quality
Vector Quantization: Pros

- Great data dependence
- Easy to learn good codebook: k-means
- Fast search
  - Offline: Pre-compute all distances between codewords
  - Online: \( d(x, y) \approx d(q(x), q(y)) \); lookup table
- Asymmetric distance
  - No need to quantize the query!
  - Online:
    - Compute all \( d(x, \text{codeword}) \)
    - \( d(x, y) \approx d(x, q(y)) \); lookup table
Vector Quantization: Cons

• Not scalable at all
  • E.g. for a standard 64-bit size code
    • Codebook size $k=2^{64}=1.8 \times 10^{19}$
  • Not possible to learn
    • Slow
      • Insufficient training data
  • Huge memory requirements for codebook and lookup tables
  • Huge cost of distance pre-computation
Product Quantization (PQ)

Key idea

- Divide vectors into $m$ subvectors (blocks)
- Vector Quantize each subvector independently
- E.g. for a standard 64-bit size code
  - 8 blocks quantized with 8 bits each
  - Sub-codebook size: $2^8 = 256$
  - Equivalent to vector quantizing with a $2^{64} = 1.8 \times 10^{19}$ codebook
- Effective codebook: codewords are the Cartesian product of block-wise codewords
Product Quantization: Distance computation

Toy example
• 2-D vectors
• Split into m=2 1-D subspaces
• Sub-codebooks of size 5
• Equivalent “product” codebook size = 25
• Assumes subspace independence

Computing distances
• Pythagoras: sum squared distances from orthogonal subspaces
• Compute squared distances in each subspace independently
• Precompute squared distances to all sub-codebook centroids
• Product quantization: 5+5 distance computations
• Vector quantization: 25 distance computations
Searching using Product Quantization

- Vector split into m subvectors: \( y \rightarrow [y_1 \ldots y_m] \)
- Subvectors are quantized separately
- Toy example: \( y = 8 \)-dim vector split into 4 subvectors of dimension 2

\( y_1 \): 2 components

\( 8^4 = 4,096 \) centroids induced

for a quantization cost equal to that of 8 centroids

- In practice: 8 bits/subquantizer (256 centroids)

Jegou et al, PAMI 2011
Searching using Product Quantization

- Estimate distances in the compressed domain
  \[ d(x, y)^2 = \sum_{i=1}^{m} d(x_i, y_i)^2 \approx \sum_{i=1}^{m} d(x_i, q_i(y_i))^2 \]

- To compute distances between query \( x \) and many codes:

  I-
  \[ \begin{array}{c}
  \begin{array}{c}
  \begin{array}{c}
  c_{1,1} \quad 1.20 \\
  c_{1,2} \quad 2.30 \\
  \vdots \\
  c_{1,8} \quad 0.34 \\
  \end{array}
  \\
  \begin{array}{c}
  c_{2,1} \quad 0.70 \\
  c_{2,2} \quad 3.01 \\
  \vdots \\
  c_{2,8} \quad 2.84 \\
  \end{array}
  \\
  \begin{array}{c}
  c_{3,1} \quad 0.15 \\
  c_{3,2} \quad 0.91 \\
  \vdots \\
  c_{3,8} \quad 1.29 \\
  \end{array}
  \\
  \begin{array}{c}
  c_{4,1} \quad 1.62 \\
  c_{4,2} \quad 0.35 \\
  \vdots \\
  c_{4,8} \quad 1.44 \\
  \end{array}
  \end{array}
  \end{array} \]
  Precompute all distances between query subvectors and centroids:
  \[ d(x_i, c_{i,j})^2 \]
  Stored in look-up tables computed per query descriptor

  II-
  For each database vector: sum the elementary square distances
  \[ m-1 \text{ additions per distance} \]
Example stats

1 Billion embeddings

Use PQ with 4 dim sub-vectors, and 1 byte per sub-vector (256 centres)

Original descriptors 128 dimension

- Memory footprint: $128 \times 4 \times 1B = 512$ GB

Product Quantization: 128 x 4 bytes -> 32 bytes per embedding

- Memory footprint: $32 \times 1B = 32$ GB

Product Quantization for vector compression,
Jegou et al., PAMI 2011
Product Quantization (PQ)

Pros

- Large reduction in memory footprint
- Large speedup in NN search (c.f. exhaustive search)
- Good performance in practice

Cons

- Slower than hashing, but usually fast enough
- Assumes subvectors are independent
  - In practice: first decorrelate subvectors (PCA)
- Assumes subvectors contain equal amount of information
  - In practice: balance information across blocks

Code: FAISS (https://github.com/facebookresearch/faiss )
ANN via approximating vectors: Comparison

Memory vs Quality
recall@100: Proportion of times the true NN is in top 100

Vector Quantization
• PQ, OPQ, CKM

Hashing
• LSH, SH, MDSH, etc

Vector Quantization beats Hashing
ANN via approximating vectors: Comparison

Memory vs Quality
recall@100: Proportion of times the true NN is in top 100

Vector Quantization
• PQ, OPQ, CKM

Hashing
• LSH, SH, MDSH, etc

Vector Quantization beats Hashing
Approximate Nearest Neighbours (ANN): Methods

Broad division

1. Approximate the vectors => do faster distance computations
   • Brute-force search, visit all points: O(n)

2. Approximate the search => do fewer distance computations
   • Do not visit all points
ANN via approximating the search: Methods

- Space partitioning
  - Hashing
  - Vector Quantization
  - K-d trees
Space partitioning

- Partition the vector space into $C$ partitions

- Query
  - Search inside the nearest $k$ partitions

- Does not decrease memory requirements
  - Can combine with quantization

- Performance
  - For $C=\text{const}$ - still $O(nd)$, just $O(nd*k/C)$
  - $k/C$: quality vs speed
Curse of dimensionality

N=1 billion, r=?

Packing hyper-spheres

1-D
r=1e-9

2-D
r=3e-5

256-D
r=0.92

\[ V_S = k_d r^d \]
\[ V_L = k_d R^d = k_d \]
\[ V_L = N V_S \]
\[ k_d = N k_d r^d \]
\[ r = \left( \frac{1}{N} \right)^{\frac{1}{d}} \]
Space partitioning revisited

- Partition the vector space into $C$ partitions

- Query
  - Search inside the nearest $k$ partitions

- Does not decrease memory requirements
  - Can combine with quantization

- Performance
  - For $C=\text{const}$ - still $O(nd)$, just $O(nd*k/C)$
  - $k/C$: quality vs speed
Space partitioning revisited

- Partition the vector space into $C$ partitions

- Query
  - Search inside the nearest $k$ partitions

- Does not decrease memory requirements
  - Can combine with quantization

- Performance
  - For $C=\text{const}$ - still $O(nd)$, just $O(nd*k/C)$
  - $k/C$: quality vs speed
ANN via approximating the search: Methods

- Space partitioning
  - Hashing
  - Vector Quantization
  - K-d trees
Hashing (for approximate search)

Recall: Each bit partitions space into two

Query

- Visit partitions in order of increasing Hamming distance of the query hash
Hashing (for approximate search): Discussion

How many partitions with Hamming distance $r$ for $b$-bits? $\binom{b}{k}$

- For $b=64$, $r=3$: $42k$
- For $b=64$, $r=4$: $635k$

Necessary solution

- Multiple partitions with smaller number of bits each
- Same vector (or its ID) is stored multiple times

Pros

- If hashing is used for quantization, partitions are directly available

Cons

- Large memory requirements for good performance
- As before: hard to get good hashing functions
ANN via approximating the search: Methods

- Space partitioning
  - Hashing
  - **Vector Quantization**
  - K-d trees
Vector Quantization (for approximate search)

Partition the space using k-means

Query
• Partition visit order: k nearest partitions obtained via NN search for cluster centres

Complexity
• Ordering partitions: $O(Cd)$
• Searching within partitions: $O(ndk/C)$
• Rule of thumb: $C \sim \sqrt{n}$
Vector Quantization (for approximate search)

Pros

• Simple and intuitive
• As before: good data dependence
• Generally good performance

Cons

• Ordering partitions can take a significant chunk of the computation budget
  • Can do ANN for this stage as well, though usually ok

Code: FAISS (https://github.com/facebookresearch/faiss)

Further reading

• Can obtain finer partitioning through PQ, see Inverted Multi-Index [Babenko & Lempitsky 2012]
ANN via approximating the search: Methods

- Space partitioning
  - Hashing
  - Vector Quantization
- K-d trees
K-d tree

A k-d tree hierarchically decomposes the vector space.

Points nearby in the space can be found (hopefully) by backtracking around the tree some small number of steps.
K-d tree

- K-d tree is a binary tree data structure for organizing a set of points in a K-dimensional space.
- Each internal node is associated with an axis aligned hyper-plane splitting its associated points into two sub-trees.
- Dimensions with high variance are chosen first.
- Position of the splitting hyper-plane is chosen as the mean/median of the projected points.
K-d tree construction

Simple 2D example

Slide credit: Anna Atramentov
K-d tree query

Slide credit: Anna Atramentov
K-d tree: Backtracking

Backtracking is necessary as the true nearest neighbor may not lie in the query cell.

But in some cases, almost all cells need to be inspected.

Figure: A. Moore
K-d tree: Backtracking

Backtracking is necessary as the true nearest neighbor may not lie in the query cell.

But in some cases, almost all cells need to be inspected.
Randomized K-d trees

- Multiple randomized trees increase the chances of finding nearby points (shared priority queue)

- How to choose the dimension to split and the splitting point?
  - Pick random dimension from high variance ones
  - Split at the mean/median
Randomized K-d trees: Discussion

Pros

• Find approximate nearest neighbor in $O(\log n)$ time, where $n$ is the number of data points
• Good ANN performance in practice for “low” dimensions (e.g. 128-D)

Cons

• Increased memory requirements: needs to store multiple (~8) trees
• Not so good in “high” dimensions

Code available online:
http://www.cs.ubc.ca/research/flann/
Approximate Nearest Neighbour search: Overview

**Approximate the vectors**: fast distances, memory savings
- Hashing
  - LSH, see ITQ, Spectral Hashing, ..
- Vector Quantization
  - *Product Quantization*, see OPQ, Cartesian K-means, ..

**Approximate the search**
- Non-exhaustive through space partitioning
- Hashing
- Vector Quantization
- (Randomized) K-d trees
- Mind the dimensionality

In real life – use a **hybrid** approach
- E.g. Google Goggles/Photos 2011: Vector Quantization + OPQ
There is more …

Inverted file indexing (Computer Vision lectures)

Beyond Euclidean distance / scalar product
  • MinHash for Jaccard similarity (set overlap)
  • Winner Take All hashing for rank correlation
  • Other kernels: Explicit feature map

Supervised methods - current and future work
  • Training representations for ANN search
    • Neural networks for hashing
      • “Supervised Hashing for Image Retrieval via Image Representation Learning” [Xia et al. 2014]
    • Neural networks for PQ
      • “SUBIC: A supervised, structured binary code for image search” [Jain et al. 2017]