Image Analysis (FMAN20)
Lecture 6, 2018
MAGNUS OSKARSSON
Image Analysis - Motivation
Overview – Machine Learning 2

• Machine Learning – supervised vs unsupervised
• More Classification
  • Nearest Neighbor
  • Logistic Regression
• Support Vector Machines
• Discriminants
• Multiclass problems
• Regression Trees
Machine Learning

• Supervised learning - classification
  • Training data consists of many pairs \((x_1,y_1), \ldots, (x_n,y_n)\)
  • Here \(x_i\) are examples of input and \(y_i\) are the corresponding correct output
  • The estimated model is used to classify future data \(f(x)=y\)

• Unsupervised learning - clustering
  • Training data consists of input data only \(x_1, \ldots, x_n\)
  • After training, the examples are clustered in groups
  • Also future data \(x\) can be assigned to groups
<table>
<thead>
<tr>
<th>Input</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Pear" /></td>
<td>Pear</td>
</tr>
<tr>
<td><img src="image2.png" alt="Apple" /></td>
<td>Apple</td>
</tr>
<tr>
<td><img src="image3.png" alt="Pear" /></td>
<td>Pear</td>
</tr>
<tr>
<td><img src="image4.png" alt="Pear" /></td>
<td>Pear</td>
</tr>
<tr>
<td><img src="image5.png" alt="Apple" /></td>
<td>Apple</td>
</tr>
<tr>
<td><img src="image6.png" alt="Apple" /></td>
<td>Apple</td>
</tr>
</tbody>
</table>

**Training Dataset**

Input: ![Apple](image7.png)  
Label: ?
• Machine Learning typically has two phases
  • Phase 1 – Training
    • A training dataset is used to estimate model parameters. Store these parameters. Code usually assumes that input are vectors
  • Phase 2 – Prediction
    • Once the parameters have been estimated, we can use the model to classify future data
Overview – Machine Learning 2

- Machine Learning
- More Classification
  - Nearest Neighbor
  - Logistic Regression
  - Support Vector Machines
  - Discriminants
  - Multiclass problems
  - Regression Trees
Nearest Neighbor Classification

NN and K-NN

- Classify using training data \((x_i, y_i)\)
- NN: Use the label of the nearest neighbor
- KNN: Use the label of the majority of the k nearest neighbors
- Regression: Use the average of the value of the k nearest neighbors
- Easy to implement and understand
- Can use arbitrary distance functions between images
- Converges to the optimum
- Slow when using lots of data, need to store all training data, not smooth regression
Nearest Neighbor Classification
7 Nearest Neighbor Classification
7 Nearest Neighbor Classification
Nearest Neighbor Classification
NN and K-NN

- Training is easy, just store the training data $T = \{(x_1, y_1), \ldots, (x_N, y_N)\}$
- Works in any dimension
- Works for regression also: Use the average of the value of the $k$ nearest neighbors
- Easy to implement and understand
- Can use arbitrary distance functions between images
- Converges to the optimum

- Slow when using lots of data,
- Need to store all training data
- Not smooth regression
Overview – Machine Learning 2

- Machine Learning
- More Classification
  - Nearest Neighbor
  - **Logistic Regression**
  - Support Vector Machines
  - Discriminants
  - Multiclass problems
  - Regression Trees
Logistic regression

• Motivation

• In the end we are only interested in the posterior distribution

\[ P(Y = y | X = x) \]

• Why not estimate this instead

• Skip the step of estimating the measurement densities

• Details far away from the transition points are uninteresting (perhaps)

• Notice that the posterior looks like a smoothed step function
Logistic regression

• \( z = \) simple function of \( x \), e.g. Linear \( z = w^T x + b \)

• Output \( y \) = smooth threshold of \( z \), for example

\[
s(z) = \frac{1}{1 + e^{-z}}
\]

• Notice that \( s(z) \) looks like a typical \( P(Y=1 \mid x) \) function

\[
x \in \mathbb{R}^d, w \in \mathbb{R}^d, b \in \mathbb{R}, f(x) = s(w^T x + b)
\]

\[
P(Y = 1 \mid x) = \frac{1}{1 + e^{-z}}
\]
Derivation

• Estimate parameters

\[ P(Y = 1|x) = \frac{1}{1 + e^{-z}} \]

\[ P(Y = -1|x) = 1 - \frac{1}{1 + e^{-z}} = \frac{e^{-z}}{1 + e^{-z}} = \frac{1}{e^z + 1} \]

• For both cases we have

\[ P(Y = y|x) = \frac{1}{1 + e^{-yz}} \]

• Calculate likelihood for training data

\[ T = (x_1, y_1), \ldots, (x_n, y_n) \]
Estimate parameters

- Parameters \( \theta = (w, b) \)

\[
P(Y = y|x) = \frac{1}{1 + e^{-yz}}
\]

\( T = (x_1, y_1), \ldots, (x_n, y_n) \)

\[
\log(P) = \log(\prod_i P(Y = y_i|x_i, \theta))
\]

\[
\sum_i \log\left( \frac{1}{1 + e^{y_i(w^T x_i + b)}} \right)
\]

\[
\min_w \frac{1}{2} w^T w + C \sum_{i=1}^l \log(1 + e^{-y_i w^T x_i})
\]
Logistic regression

- Linear logistic regression
- Estimate the posterior $P(Y = y|X = x)$
- As linear function followed by standard logistic function
  $$s(w^T x + b)$$
- Convex optimization problem

Standard logistic function
$$s(z) = \frac{1}{1 + e^{-z}}$$
Overview – Machine Learning 2

- Machine Learning
- More Classification
  - Nearest Neighbor
  - Logistic Regression
- **Support Vector Machines**
- Discriminants
- Multiclass problems
- Regression Trees
Support Vector Machines

- Formulation:
  \[
  \min_{w,b} \quad \frac{1}{2} \|w\|^2 \\
  \text{such that} \quad y_i(w^T x_i + b) \geq 1
  \]
- Quadratic program with linear constraints
Solving the Optimization Problem

\[
\begin{align*}
\min_{w,b} & \quad \frac{1}{2} \|w\|^2 \\
\text{s.t.} & \quad y_i(w^T x_i + b) \geq 1
\end{align*}
\]

The Lagrangian needs to be minimized w.r.t. \( w, b \), and maximized w.r.t. \( \alpha_i \)
Solving the Optimization Problem

\[
\min_{\mathbf{w}, b} \max_{\alpha} L_p(\mathbf{w}, b, \alpha_i) = \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^{n} \alpha_i \left( y_i (\mathbf{w}^T \mathbf{x}_i + b) - 1 \right)
\]

s.t. \( \alpha_i \geq 0 \)

\[
\frac{\partial L_p}{\partial \mathbf{w}} = 0 \quad \Rightarrow \quad \mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i
\]

\[
\frac{\partial L_p}{\partial b} = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} \alpha_i y_i = 0
\]

Solution is an expansion in terms of training examples

Due to strict convexity, \( \mathbf{w} \) is unique although \( \alpha_i \)'s need not be
Solving the Optimization Problem

\[
\min_{w,b} \max_{\alpha} \quad L_p(w, b, \alpha_i) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i \left( y_i (w^T x_i + b) - 1 \right)
\]

s.t. \quad \alpha_i \geq 0

Lagrangian Dual Problem

\[
\max_{\alpha} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j
\]

s.t. \quad \alpha_i \geq 0 \quad \text{and} \quad \sum_{i=1}^{n} \alpha_i y_i = 0
Solving the Optimization Problem

- From the KKT conditions, we know:

\[ \alpha_i \left( y_i (\mathbf{w}^T \mathbf{x}_i + b) - 1 \right) = 0 \]

- Thus, only support vectors have \( \alpha_i \neq 0 \)

- The solution has the form:

\[ \mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i = \sum_{i \in SV} \alpha_i y_i \mathbf{x}_i \]

get \( b \) from \( y_i (\mathbf{w}^T \mathbf{x}_i + b) - 1 = 0 \),

where \( \mathbf{x}_i \) is support vector \( b = \frac{1}{N_{SV}} \sum_{i=1}^{N_{SV}} (\mathbf{w}^T \mathbf{x}_i - y_i) \)

more robust estimate, average over SVs
Solving the Optimization Problem

- The linear discriminant function is:

\[ g(x) = w^T x + b = \sum_{i \in SV} \alpha_i x_i^T x + b \]

- Relies on a *dot product* between the test point \( x \) and the support vectors \( x_i \)

- Solving the optimization problem involved computing the *dot products* \( x_i x_i^T \) between all pairs of training points
`Soft Margin’ Linear Classifier

- What if data is not linear separable due to noise or outliers?

- Slack variables $\xi_i$ can be added to allow for the mis-classification of difficult or noisy data.
`Soft Margin’ Linear Classifier

- Formulation:

\[
\min_{w,b,\xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi_i
\]

such that

\[
y_i (w^T x_i + b) \geq 1 - \xi_i
\]

\[
\xi_i \geq 0
\]

- for \(0 \leq \xi \leq 1\), point is between margin and correct side of hyperplane
- for \(\xi > 1\), point is misclassified

- Parameter \(C\) can be viewed as a means to control over-fitting
  - small \(C\) allows constraints to be easily ignored: *large margin*
  - large \(C\) makes constraints hard to ignore: *narrow margin*
  - \(C = \infty\) enforces all constraints: *hard margin*
`Soft Margin’ Linear Classifier

- Formulation (Lagrangian Dual Problem)

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

such that

$$0 \leq \alpha_i \leq C$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$
`Soft Margin’ Interpretation (I)

- The constraint \( y_i (\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i \) can be written more concisely as
  \[
y_i g(\mathbf{x}_i) \geq 1 - \xi_i \iff \xi_i = \max(0, 1 - y_i g(\mathbf{x}_i))
  \]

- Hence we need to solve the learning problem
  \[
  \min_{\mathbf{w}, b} \|\mathbf{w}\|^2 + C \sum_{i=1}^{N} \max(0, 1 - y_i g(\mathbf{x}_i))
  \]
`Soft Margin’ Interpretation (II)

We need to solve the learning problem

\[ \min_{\mathbf{w}, b} \|\mathbf{w}\|^2 + C \sum_{i=1}^{N} \max(0, 1 - y_i g(x_i)) \]

- \( y_i g(x_i) > 1 \) ⇒ point is outside margin and does not contribute to loss
- \( y_i g(x_i) = 1 \) ⇒ point is on margin and does not contribute to loss (as in hard margin)
- \( y_i g(x_i) < 1 \) ⇒ point violates margin constraint and contributes to loss
SVM uses Hinge Loss

Can be viewed as an approximation to the 0-1 loss
Non-linear SVMs

• Datasets that are linearly separable with noise work out great:

• But what are we going to do if the dataset is just too hard?

• How about... mapping data to a higher-dimensional space:
Non-linear SVMs: Feature Space

General idea: the original input space can be mapped to some higher-dimensional feature space where the training set is separable

\[ \Phi: \mathbf{x} \rightarrow \phi(\mathbf{x}) \]
How to Use the Feature Space?

- The feature point $z = \phi(x)$ corresponding to an input point $x$ is called the image (or the lifting) of $x$; the input point $x$, if any, corresponding to a given feature vector $z$ is called the pre-image of $z$.

- The naive way to use a feature space is to explicitly compute the image of every training and testing point, and run algorithm fully in feature space.

- Two potential problems:
  - The feature space may be very high dimensional or infinite dimensional, so direct (explicit) calculations in such feature space may not be practical, or even possible.
  - We may sometimes want to map back an answer from feature space to the input space. This is called the pre-image problem. For some feature maps (kernels), analytical expressions are available, but in most other cases some form of (local) optimization may be necessary.
Nonlinear SVMs: The Kernel Trick

• With this mapping, our discriminant function is now:

\[ g(x) = w^T \phi(x) + b = \sum_{i \in SV} \alpha_i \phi(x_i)^T \phi(x) + b \]

• No need to know this mapping explicitly, because we only use the *dot product* of feature vectors both in training and in testing

• A *kernel function* is defined as a function that corresponds to a dot product of two feature vectors in some expanded feature space:

\[ K(x_i, x_j) \equiv \phi(x_i)^T \phi(x_j) \]
Positive Definite Kernels

- **Gram Matrix.** Given a function $k: X^2 \rightarrow \mathbb{R}$ (or $\mathbb{C}$), and patterns $x_1, \ldots, x_m \in X$, the $m \times m$ matrix $K$ with elements $K_{ij} := k(x_i, x_j)$ is called the Gram matrix (or kernel matrix) of $k$ w.r.t $x_1, \ldots, x_m$.

- **Positive definite kernel.** A complex $m \times m$ matrix $K$ satisfying $\sum_{ij} c_i \overline{c_j} K_{ij} \geq 0$, $\forall c_i \in \mathbb{C}$ is called positive definite. Similarly, a real symmetric $m \times m$ matrix $K$ satisfying the above for all $c_i \in \mathbb{R}$ is called positive definite.

positive definite kernels $\equiv$ Mercer kernels $\equiv$ reproducing kernels $\equiv$ admissible kernels $\equiv$ support vector kernels $\equiv$ covariance functions
Examples of Kernels

Examples of commonly-used kernel functions:

- Linear kernel: \( K(x_i, x_j) = x_i^T x_j \)

- Polynomial kernel: \( K(x_i, x_j) = (1 + x_i^T x_j)^p \)

- Gaussian (Radial-Basis Function (RBF)) kernel:
  \[
  K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)
  \]

- Sigmoid:
  \[
  K(x_i, x_j) = \tanh(\beta_0 x_i^T x_j + \beta_1)
  \]
Generality of Kernel Trick

- Given an algorithm expressed in terms of a positive-definite kernel \( k \), we can construct an alternative algorithm by replacing \( k \) with another positive-definite kernel \( \tilde{k} \).
- This is not limited to only cases when \( k \) is a dot product in the input domain.
- Any algorithm that only depends on dot products (i.e. is rotationally invariant) can be kernelized.
- Kernels are defined on general sets (rather than just dot product spaces!) and their use leads to an embedding of general data types in linear spaces.
Nonlinear SVM: Optimization

- Formulation (Lagrangian Dual Problem)

\[
\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j)
\]

such that

\[
0 \leq \alpha_i \leq C
\]

\[
\sum_{i=1}^{n} \alpha_i y_i = 0
\]

- The solution of the discriminant function is

\[
g(x) = \sum_{i \in SV} \alpha_i K(x_i, x) + b
\]
Support Vector Machine: Algorithm

1. Choose a kernel function

2. Choose a value for $C$

3. Solve the quadratic programming problem
   (many software packages available, e.g. libsvm)

4. Construct the discriminant function from the support vectors
Support Vector Machines

- SVM Applet demo
- Other demos
  - [http://cs.stanford.edu/people/karpathy/convnetjs/](http://cs.stanford.edu/people/karpathy/convnetjs/)
Properties of Kernels

- Kernels are symmetric in their arguments:
  \[ K(x_1, x_2) = K(x_2, x_1) \]

- They are positive valued for any inputs: \( K(x_1, x_2) \geq 0 \)

- The Cauchy-Schwartz inequality holds:
  \[ K^2(x_1, x_2) \leq K(x_1, x_1)K(x_2, x_2) \]

- Technically, to use a function as a kernel, it must satisfy Mercer’s conditions for a positive-definite operator

- The intuition is easy to grasp for finite spaces
  - Discretize \( x \) space as densely as desired into buckets \( x_i \)
  - Between each two cells \( x_i, x_j \), compute the kernel function, and write these values as a (symmetric) matrix \( M_{ij} = K(x_i, x_j) \)
  - If the matrix is positive definite, the kernel is OK
Kernel Closure Rules

Very useful for designing new kernels from existing kernels

– The sum of any two kernels is a kernel

– The product of any two kernels is a kernel

– A kernel plus a constant is a kernel

– A scalar times a kernel is a kernel
Support Vector Machine Detector

Training set

Multi-scale search

Test image

descriptors

training

Support vector machine

results
SVM – Pedestrian detection
Overview – Machine Learning 2

• Machine Learning
• More Classification
  • Nearest Neighbor
  • Logistic Regression
• Support Vector Machines
• Discriminants
• Multiclass problems
• Regression Trees
Multi-Class Decision Problems

• The goal in classification is to take an input vector \( \mathbf{x} \) and to assign it to one of \( K \) discrete classes \( C_k \) where \( k = 1, \ldots, K \).

• The classes are taken to be disjoint, so that each input is assigned to one and only one class (N.B. This does not necessarily mean that data is separable!)

• The input space is thereby divided into decision regions whose boundaries are called decision boundaries or decision surfaces.

• We consider linear models for classification, by which we mean that the decision surfaces are linear functions of the input vector \( \mathbf{x} \) and hence are defined by \( (D - 1) \) — dimensional hyperplanes within the \( D \) — dimensional input space.

• Data sets whose classes can be separated exactly by linear decision surfaces are said to be linearly separable.
Linear Classification

- Classification is intrinsically non-linear because of the training constraints that **place non-identical inputs in the same class**
- Differences in the input vector sometimes causes 0 change in the answer
- The **presence of more than two classes** makes formulation harder (see next)
- Linear classification means that the adaptive part $\mathbf{w}$ is linear
  - The adaptive part is cascaded with a fixed non-linearity $f$
  - It may also be preceded by a fixed non-linearity $\phi$ when nonlinear basis functions are used

$$y(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) + w_0, \quad c = f(y(\mathbf{x}))$$

Since the decision boundary is $y(\mathbf{x}) = \text{constant}$, for $\phi(\mathbf{x}) = \mathbf{x}$, the decision surfaces are linear functions of $\mathbf{x}$ even if $f$ is non-linear!
**Approach 1**: Discriminant Function

- Use discriminant functions directly, and do not compute probabilities

- Convert the input vector into one or more real values so that a simple process (threshold, or majority vote) can be applied to assign the input to the class

- The real values should be chosen to maximize the useable information about the class label present in the real value

- Given discriminant functions $f_1(x), \ldots, f_K(x)$
  Classify $x$ as class $C_k$, iff $f_k(x) > f_j(x), \forall j \neq k$
Approach 2: Class-conditional Probabilities

- Infer conditional class probabilities $p(C_k | x)$
- Use conditional distribution to make optimal decisions, e.g. by minimizing some loss function

- Example, 2 classes

\[
p(C_1 | x) = \sigma(w^T \phi), \quad p(C_2 | x) = 1 - p(C_1 | x)
\]

where \(\sigma(a) = \frac{1}{1 + \exp(-a)}\)
**Approach 3: Class Generative Model**

- Compare the probability of the input under separate, class-specific, generative models.
- Model both the class conditional densities $p(x|C_k)$, and the prior class probabilities $p(C_k)$.
- Compute posterior using Bayes’ theorem:

  $$ p(C_k|x) = \frac{p(x|C_k)p(C_k)}{p(x)} = \frac{p(x|C_k)p(C_k)}{\sum_j p(x|C_j)p(C_j)} $$

- Example: fit a multivariate Gaussian to the input vectors corresponding to each class, model class prior probabilities by training data frequency counts, and see which Gaussian makes a test data vector most probable using Bayes’ theorem.
2-class case: The decision surface in data-space for the linear discriminant function

\[ y(x) = w^T x + w_0 \]

\( w \) is orthogonal to any vector which lies on the decision surface, \( y(x) = 0 \)

\( w \) controls the orientation of the decision surface

\[ y(x_{\perp}) = w^T x_{\perp} + w_0 = 0 \]

\[ x = x_{\perp} + r \frac{w}{||w||} \]

\[ r = \frac{y(x)}{||w||} \]
Represent Target Values: Binary vs. Multiclass

- **Two classes (N=2):** typically use a single real valued output that has target values of 1 for the positive class and 0 (or -1) for the negative class
  - For probabilistic class labels, the target can be the probability of the positive class and the output of the model can be the probability the model assigns to the positive class

- **For the multiclass (N>2),** we use a vector of \( N \) target values containing a single 1 for the correct class and zeros elsewhere
  - For probabilistic labels we can then use a vector of class probabilities as the target vector
Discriminant Functions for Multiclass

- One possibility is to use $N$ binary (2-way) discriminants
  - Each function separates one class from the rest.

- Another possibility is to use $\frac{N(N-1)}{2}$ binary (2-way) discriminants
  - Each function discriminates between two specific classes. We have 1 discriminant for each class pair. For decision, we can take a majority vote.

- Both methods have ambiguities
Problems with Multi-class Discriminant Functions Constructed from Binary Classifiers

If we base our decision on binary classifiers, we can encounter ambiguities
Simple Solution

Use $N$ discriminant functions, $y_i, y_j, y_k, \ldots$, and take the max over their response

- Consider linear discriminants $y$
  
  $$y_k(x) = w_k^T x + w_{k0}$$

- The decision boundary between class $k$ and $j$ is given by the $D - 1$ hyperplane
  
  $$(w_k^T - w_j^T)x + (w_{k0} - w_{j0}) = 0$$
  (analogous geometrical properties as in the two class case apply)

In this linear case the decision regions are convex (max will give consistent results)

$$x_A, x_B \in R_k, \hat{x} = \lambda x_A + (1 - \lambda) x_B, 0 \leq \lambda \leq 1$$

From the linearity of $y \Rightarrow y_k(\hat{x}) = \lambda y_k(x_A) + (1 - \lambda)y_k(x_B)$

But $y_k(x_A) > y_j(x_A)$ and $y_k(x_B) > y_j(x_B) \quad \forall j \neq k \Rightarrow y_k(\hat{x}) > y_j(\hat{x})$

$\Rightarrow \hat{x}$ also lies inside $R_k$

Hence $R_k$ is convex
Fisher’s Linear Discriminant

• We can view classification in terms of dimensionality reduction
• A simple linear discriminant function is a projection of the $D$-dimensional data $\mathbf{x}$ down to 1 dimension

  \[
  \text{Project: } y = \mathbf{w}^T \mathbf{x};
  \]

  \[
  \text{Classify: if } y \geq -\mathbf{w}_0 \text{ then } C_1 \text{ else } C_2
  \]

• However projection results in loss of information. Classes well separated in the original input space may strongly overlap in 1d

• However, by adjusting the projection (weight vector $\mathbf{w}$) we can aim at the best separation among classes. But what do we mean by best separation?
Fisher’s View of Class Separation (I)

- Consider a two-class problem. There are $N_1$ points of class $C_1$ and $N_2$ of class $C_2$.
- The simplest measure of class separation when projected onto $\mathbf{w}$ is the one between the projected class means. This suggests choosing $\mathbf{w}$ so to maximize

$$m_2 - m_1 = \mathbf{w}^T(m_2 - m_1), \quad m_k = \mathbf{w}^T \mathbf{m}_k, \quad m_1 = \frac{1}{N_1} \sum_{n \in C_1} \mathbf{x}_n, \quad m_2 = \frac{1}{N_2} \sum_{n \in C_2} \mathbf{x}_n$$

- This can be made arbitrarily large by increasing $||\mathbf{w}||$. We could handle this by imposing unit norm constraints using Lagrange multipliers. We get

$$\max_{\mathbf{w}} \mathbf{w}^T(m_2 - m_1), \text{ s. th. } ||\mathbf{w}|| = 1 \Rightarrow \mathbf{w} \propto m_2 - m_1$$

- However, still, if the main direction of variance in each class is not orthogonal to the direction between means, we will not get good separation (see next slide). This is often due to strongly non-diagonal class distribution covariances.
Advantage of using Fisher’s Criterion

Although well separated in the original 2d space, when projected onto the 1d (line) joining their means, the classes are not well separated.

Fisher chooses a direction that makes the projected classes much tighter, even though their projected means are less far apart.
Fisher’s View of Class Separation (II)

Fisher: maximize a function that gives a large separation between the projected class means, while also giving a small variance within each class, thereby minimizing class overlap

– Choose direction maximizing the ratio of between class variance to within class variance
– This is the direction in which the projected points contain the most information about class membership (under Gaussian assumptions)
Fisher’s Linear Discriminant

- We seek a linear transformation that is best for discrimination
  \[ y = \mathbf{w}^T \mathbf{x} \]
- Projection onto the vector separating the class seems right, initially
  \[ \mathbf{w} \propto \mathbf{m}_2 - \mathbf{m}_1 \]
- But we also want small variance within each class
  \[ s_k^2 = \sum_{n \in C_k} (y_n - m_k)^2, \quad m_k = \mathbf{w}^T \mathbf{m}_k \]
- Fisher’s objective function
  \[ J(\mathbf{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2} \]
  - Between class
  - Within class
Fisher’s Linear Discriminant Derivations

\[ J(w) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2} = \frac{w^T S_B w}{w^T S_W w} \]

where

\[ S_B = (m_2 - m_1)(m_2 - m_1)^T \]
\[ S_W = \sum_{n \in C_1} (x_n - m_1)(x_n - m_1)^T + \sum_{n \in C_2} (x_n - m_2)(x_n - m_2)^T \]

\[ \frac{d J(w)}{dw} = 0 \Rightarrow (w^T S_B w)S_W w = (w^T S_W w)S_B w \]

Optimal solution: \[ w \propto S_W^{-1} (m_2 - m_1) \propto (m_2 - m_1) \]

The above result is known as Fischer’s linear discriminant. Strictly it is not a discriminant, but rather a direction of projection that can be used for classification in conjunction with a decision (e.g. threshold) operation.
Fischer’s Linear Discriminant Computation

However, the objective $J(w)$ is invariant to rescaling $w \rightarrow \alpha w$. We can chose the denominator to be unity. We can then minimize

$$\min_w -\frac{1}{2} w^T S_B w$$

$$w^T S_W w = 1$$

This corresponds to the (primal) Lagrangian

$$L_P = -\frac{1}{2} w^T S_B w + \frac{1}{2} \lambda( w^T S_W w - 1)$$

From the KKT conditions

$$S_B w = \lambda S_W w \Rightarrow S_W^{-1} S_B w = \lambda w$$

Generalized eigenvalue problem, as $S_W^{-1} S_B$ not symmetric
Fischer’s Linear Discriminant Computation

\[ S_W^{-1} S_B w = \lambda w \]

- Given that \( S_B \) is symmetric positive definite, we can write
  
  \[ S_B = S_B^{1/2} S_B^{1/2} \]

  where \( S_B = U \Sigma U^T \), \( S_B^{1/2} = U \Sigma^{1/2} U^T \)

- Defining \( v = S_B^{1/2} w \), we get
  
  \[ S_B^{1/2} S_W^{-1} S_B^{1/2} v = \lambda v \]

- We have to solve a regular eigenvalue problem for a symmetric, positive definite matrix \( S_B^{1/2} S_W^{-1} S_B^{1/2} \)
  - We can find solutions \( \lambda_k \) and \( v_k \) corresponding to \( S_B^{1/2} w \)

- \textit{Which eigenvector and eigenvalue should we choose? The largest! Why?}

- Transforming to dual

\[ w^T S_B w = 1, \quad w^T S_W w = \frac{1}{\lambda_k} \Rightarrow L_D = \text{const.} + \frac{1}{2} \lambda_k \]

(need to maximize over \( \lambda \))
The Logistic Sigmoid (due to S-shape)

- This is also called a squashing function because it maps the entire real axis into a finite interval.

- For classification, the output $a$ is a smooth function of the inputs and the weights $\mathbf{w}$: $a = \mathbf{w}^T \mathbf{x} + w_0$

- Properties:
  
  $$\sigma(-a) = 1 - \sigma(a), \quad a = \ln\left(\frac{\sigma}{1 - \sigma}\right)$$

  logit function

  $$y = \sigma(a) = \frac{1}{1 + e^{-a}}$$

  $$\frac{\partial a}{\partial w_i} = x_i \quad \frac{\partial a}{\partial x_i} = w_i$$

  $$\frac{dy}{da} = y (1 - y)$$
Probabilistic Generative Models

• Use a class prior and a separate generative model of the input vectors for each class, and compute which model makes a test input vector most probable
• The posterior probability of class 1 is given by:

\[
p(C_1 | x) = \frac{p(C_1)p(x | C_1)}{p(C_1)p(x | C_1) + p(C_2)p(x | C_2)} = \frac{1}{1 + e^{-a}}
\]

where \( a = \ln \frac{p(C_1)p(x | C_1)}{p(C_2)p(x | C_2)} = \ln \frac{p(C_1 | x)}{1 - p(C_1 | x)} \)

z is called the logit and is given by the log odds
Multiclass Model (Softmax)

\[ p(C_k | x) = \frac{p(x|C_k)p(C_k)}{p(x)} = \frac{p(x|C_k)p(C_k)}{\sum_j p(x|C_j)p(C_j)} = \frac{\exp(a_k)}{\sum_j \exp(a_j)} \]

where \( a_k = \ln p(x|C_k)p(C_k) \)

- This is known as the normalized exponential
- Can be viewed as a multiclass generalization of the logistic sigmoid
- It is also called a softmax function (it is a smoothed version of `max’)

if \( a_k \gg a_j \forall j \neq k \), then \( p(C_k|x) \approx 1 \) and \( p(C_j|x) \approx 0 \)
Gaussian Class-Conditionals

- Assume that the input vectors for each class are from a Gaussian distribution, and all classes have the same covariance matrix. The class conditionals are

\[ p(x \mid C_k) = \frac{1}{Z} \exp\left\{ -\frac{1}{2} (x - \mu_k)^T \Sigma^{-1} (x - \mu_k) \right\} \]

(normalizer) \quad ((same) inverse covariance matrix)

- For two classes, \( C_1 \) and \( C_2 \), the posterior turns out to be a logistic

\[ p(C_1 \mid x) = \sigma(w^T x + w_0) \]

\[ w = \Sigma^{-1} (\mu_1 - \mu_2) \]

(Quadratic terms in x canceled due to common covariance)

\[ w_0 = -\frac{1}{2} \mu_1^T \Sigma^{-1} \mu_1 + \frac{1}{2} \mu_2^T \Sigma^{-1} \mu_2 + \ln \frac{p(C_1)}{p(C_2)} \]
Interpretation of Decision Boundaries

\[ p(C_1 \mid x) = \sigma(w^T x + w_0) \]
\[ w = \Sigma^{-1}(\mu_1 - \mu_2) \]
\[ w_0 = -\frac{1}{2} \mu_1^T \Sigma^{-1} \mu_1 + \frac{1}{2} \mu_2^T \Sigma^{-1} \mu_2 + \ln \frac{p(C_1)}{p(C_2)} \]

- Quadratic terms canceled due to common covariance
- The sigmoid takes a linear function of \( x \) as argument
- The decision boundaries correspond to surfaces along which the posteriors \( p(C_k \mid x) \) are constant, so they will be given by linear functions of \( x \). Thus, decision boundaries are linear functions in input space
- The prior probabilities \( p(C_k) \) enter only through the bias parameter \( w_0 \), so changes in priors have the effect of making parallel shifts of the decision boundary (more generally of the parallel contours of constant posterior probability)
A picture of the two Gaussian models and the resulting posterior for the red class

$p(x|C_1), p(x|C_2)$

$p(C_1|x)$

The logistic sigmoid in the right-hand plot is coloured using a proportion of red tone given by $p(C_1|x)$ and a proportion of blue tone given by $p(C_2|x) = 1 - p(C_1|x)$. 
Class posteriors when covariance matrices are different for different classes

$p(x|C_k)$

$p(C_k|x)$

The decision surface is planar when the covariance matrices are the same and quadratic when they are not.
Overview – Machine Learning 2

• Machine Learning
• More Classification
  • Nearest Neighbor
  • Logistic Regression
• Support Vector Machines
• Discriminants
• Multiclass problems
• Regression Trees
Regression trees
Decision trees advantages

- Simple to understand and interpret
- Requires little preparation
- Can handle both continuous and discrete data
- ‘white box’ model. You can easily explain a decision afterwards
- Robust
- Performs well with large datasets
Decision tree limitations

- Optimal learning is NP-complete (use heuristics)
- Problems with over fitting
Regression trees learning

- Try each variable
- Try each threshold
- Calculate score e.g.
  - Entropy
    \[ I_E(f) = - \sum_{i=1}^{m} f_i \log_2 f_i \]
  - Gini Impurity
    \[ I_G(f) = \sum_{i=1}^{m} f_i (1 - f_i) = \sum_{i=1}^{m} (f_i - f_i^2) = \sum_{i=1}^{m} f_i - \sum_{i=1}^{m} f_i^2 = 1 - \sum_{i=1}^{m} f_i^2 \]
Regression trees learning

- \( \log_2(6) = 2.58 \) (bits)
- \( f = [0.24, 0.23, 0.11, 0.13, 0.21, 0.08] \)
- \( I(f) = 2.48 \) (bits)
- Try each threshold
- Calculate score e.g.
  - Entropy

\[
I_E(f) = - \sum_{i=1}^{m} f_i \log_2 f_i
\]
Summary

- Machine Learning
- Unsupervised Learning -
- Supervised Learning -
- More Classification
  - Nearest Neighbor
  - Logistic Regression
  - Support Vector Machines
  - Discriminants
  - Multiclass problems
  - Regression Trees
Master’s Thesis Suggestion
Collaborative SLAM
How can many cars map the world