Foundation of Intelligent Systems, Part I

Classification

mcuturi@i.kyoto-u.ac.jp
Last Lecture: Regression

• Mentioned the Maximum Likelihood perspective on LS-regression

\[
\log \mathcal{L}(a, b) = C - \frac{1}{2\sigma^2} \sum_{j=1}^{N} \| y_j - (a^T x_j + b) \|^2.
\]

• Provided a geometric perspective on LS regression through projections

<table>
<thead>
<tr>
<th>Least Squares Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \updownarrow )</td>
</tr>
<tr>
<td>Projecting the vector of observed predicted variable in ( \text{span}{\text{vectors of observed predictor variables + constant vector}} )</td>
</tr>
</tbody>
</table>

• Many issues with LS regression... Hence advanced regression techniques
  
  ○ Ridge Regression  
  ○ Subset selection  
  ○ Lasso

• we will talk about these in 3 lectures when discussing sparsity.
Today

- Classification, differences with regression
- Binary classification
- Linear classification algorithms
  - Logistic Regression
  - Ideally, Linear Discriminant Analysis, but no time.
  - Perceptron rule
  - Support Vector Machine
- Once this is done, we will move on to more theory in next lecture about statistical learning theory.
Classification
Many observations of the same data type, with label

- we still consider a database \( \{x_1, \cdots, x_N\} \),

- each datapoint \( x_j \) is represented as a vector of features \( x_j = \begin{bmatrix} x_{1,j} \\ x_{2,j} \\ \vdots \\ x_{d,j} \end{bmatrix} \)

- To each observation is associated a label \( y_j \)...
  - If \( y_j \in \mathbb{R} \), we have **regression**
  - If \( y_j \in S \) where \( S \) is a finite set, **multiclass classification**.
  - If \( S \) only has two elements, **binary classification**.
Examples

Multiclass Classification

- Classify images of fruits into fruit category
- Classify images of handwritten digits into digits from 0 to 9
- Classify musical tunes, books, movies into genres
- Classify proteins into functional classes
Examples

Binary Classification

• Using elementary measurements, guess if someone has or not a disease that is
  ○ difficult to detect at an early stage
  ○ difficult to measure directly (fetus)

• Classify chemical compounds into toxic / nontoxic

• Classify a passenger as suspect/not suspect

• Classify body tumor as begin/malign to detect cancer

• etc.
Why use a new name?

Solve a **classification** problem ⇔ **build a function** $f : \mathbb{R}^d \rightarrow S$
Why use a new name?

Solve a classification problem $\iff$ build a function $f : \mathbb{R}^d \rightarrow S$

To do so, we need to evaluate the accuracy of a function $f$. 
Why use a new name?

Solve a classification problem \(\iff\) build a function \(f : \mathbb{R}^d \to S\)

To do so, we need to evaluate the accuracy of a function \(f\),

Namely, for each \(j\), can we measure whether \(f(x_j) \approx y_j\)?
Why use a new name?

Solve a **classification** problem ⇔ **build a function** $f : \mathbb{R}^d \rightarrow S$

for each $j$, can we measure whether $f(x_j) \approx y_j$?

**In conventional regression - linear regression**

- We have used consistently $\sum_{j=1}^{N} \| f(x_j) - y_j \|^2$ to select a good $f$.
- $\mathbb{R}$ is a **metric** space... $\| 37.354 \text{ JPY} - 36.000 \text{ JPY} \| = 1354$
  - sense of closeness between possible answers
- $\mathbb{R}$ is a totally **ordered** set... $36.000 \text{ JPY} < 37.354 \text{ JPY}$
  - notion of total hierarchy between possible answers
Why use a new name?

Solve a classification problem ⇔ build a function $f : \mathbb{R}^d \rightarrow S$
for each $j$, can we measure whether $f(x_j) \approx y_j$?

In discrete labels in classification

- No distance nor order is available on $S$
  - No order for musical genres $\text{jazz} > \text{bossa-nova}$?
  - No distance between fruits $\|\text{kiwi} - \text{banana}\|$?
- This creates challenges to quantify how $f(x_j)$ is close to $y_j$
Digits recognition

- Use a database such as

\[
\begin{array}{ccccccccccccccc}
2 & 6 & 0 & 1 & 9 & 2 & 7 & 1 & 4 & 0
\end{array}
\]

paired with the corresponding labels,

\[
(2, 6, 0, 1, 9, 2, 7, 1, 4, 0, 5, 4, 3, 0, 8, 4, 3, 9, 4, 7).
\]

to build an **automated recognition system** for handwritten digits.

- useful for post office, check recognition, tax office, etc..
Labels are usually unordered and without a metric

- The set of labels is $S = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$
- Yet there is no distance/order in $S$ for this task.
- Suppose the image given to the recognition system is

\[ \begin{array}{c}
  6 \\
\end{array} \]

- Although $|5 - 6| < |0 - 6|$, the answer 5 is not better than 0.
Sometimes discrete labels can be given with a metric

- Suppose the task is to guess the rating in stars of a movie

<table>
<thead>
<tr>
<th>DVDs in Your Queue</th>
<th>Star Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>Priority</td>
<td>Movie Title</td>
</tr>
<tr>
<td>1</td>
<td>We and You and Everyone We Know</td>
</tr>
<tr>
<td>2</td>
<td>Weeds and Weeds</td>
</tr>
<tr>
<td>3</td>
<td>Arrested Development: Season 2</td>
</tr>
<tr>
<td>4</td>
<td>Arrested Development: Season 2</td>
</tr>
<tr>
<td>5</td>
<td>Arrested Development: Season 2</td>
</tr>
<tr>
<td>6</td>
<td>Look of Silence</td>
</tr>
<tr>
<td>7</td>
<td>The United States of Leland</td>
</tr>
<tr>
<td>8</td>
<td>Danny O’Day: Munkey’s Cup</td>
</tr>
<tr>
<td>9</td>
<td>Eternal Sunshine of the Spotless Mind</td>
</tr>
<tr>
<td>10</td>
<td>Lost in Translation</td>
</tr>
<tr>
<td>11</td>
<td>Not with a Shot Glass</td>
</tr>
<tr>
<td>12</td>
<td>Sean Penn</td>
</tr>
<tr>
<td>13</td>
<td>Elephant</td>
</tr>
<tr>
<td>14</td>
<td>The Red Violin</td>
</tr>
<tr>
<td>15</td>
<td>Speak</td>
</tr>
</tbody>
</table>

- User inputs are in $S = \{1, 2, 3, 4, 5\}$

- In this case standard regression techniques may be applied because,
  - the natural metric $\|5 - 3\|$ works
  - linear regression works because the order is also valid.
  - the final user does not mind getting fractional predictions (e.g. 3.85)
Binary Classification

\[
\text{card } S = 2. \\
\]

Usually \( S = \{0, 1\} \) or \( S = \{-1, 1\} \) or \( S = \{-, +\} \) or \( S = \{Y, N\} \)
Data

- The **Data** we have: a bunch of vectors $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \cdots , \mathbf{x}_N$.
- Ideally, to infer a “yes/no” rule, we need the **correct answer** for each vector.
- We consider thus a set of **pairs of** (vector,bit)

$$
\text{“training set”} = \left\{ \left( \mathbf{x}_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{id} \end{bmatrix} \in \mathbb{R}^d, \ y_i \in \{0, 1\} \right) \right\}_{i=1..N}
$$

- For illustration purposes **only** we will consider **vectors in the plane**, $d = 2$.
- Points are easier to represent in 2 dimensions than in 20.000...
- The ideas for $d \gg 3$ are **exactly the same**.
What is a classification rule?
Binary Classification Separation Surfaces for Vectors

Classification rule = a partition of $\mathbb{R}^d$ into two sets
This partition is usually interpreted as the level set of function on $\mathbb{R}^d$. 
Binary Classification Separation Surfaces for Vectors

Typically, \( \{ \mathbf{x} \in \mathbb{R}^d | f(\mathbf{x}) > 0 \} \) and \( \{ \mathbf{x} \in \mathbb{R}^d | f(\mathbf{x}) \leq 0 \} \)
Classification Separation Surfaces for Vectors

Can be defined by a single surface, *e.g.* a curved line
Even more simple: using straight lines and halfspaces.
Linear Classifiers

- **Straight lines** (hyperplanes when \( d > 2 \)) are the simplest type of classifiers.
- A hyperplane \( H_{c,b} \) is a set in \( \mathbb{R}^d \) defined by
  - a normal vector \( c \in \mathbb{R}^d \)
  - a constant \( b \in \mathbb{R} \).

\[
H_{c,b} = \{ x \in \mathbb{R}^d \mid c^T x = b \}
\]

- Letting \( b \) vary we can “slide” the hyperplane across \( \mathbb{R}^d \)
Linear Classifiers

- Exactly like lines in the plane, hypersurfaces divide $\mathbb{R}^d$ into two halfspaces,

$$\{x \in \mathbb{R}^d \mid c^T x < b\} \cup \{x \in \mathbb{R}^d \mid c^T x \geq b\} = \mathbb{R}^d$$

- Linear classifiers attribute the “yes” and “no” answers given arbitrary $c$ and $b$.

- Assuming we only look at halfspaces for the decision surface...

  ...how to choose the “best” $(c^*, b^*)$ given a training sample?
Linear Classifiers

- This specific question, 

  \[
  \{ (x_i \in \mathbb{R}^d, y_i \in \{0, 1\}) \}_{i=1..N} \xrightarrow{???} \text{“best” } c^*, b^*
  \]

  has different answers. Depends on the meaning of “best”?:

- **Linear Discriminant Analysis** (or Fisher’s Linear Discriminant);
- **Logistic regression** maximum likelihood estimation;
- **Perceptron**, a one-layer neural network;
- **Support Vector Machine**, the result of a convex program
- *etc.*
Given two sets of points...
Classification Separation Surfaces for Vectors

It is sometimes possible to separate them perfectly
Each choice might look equivalently good on the training set, but it will have obvious impact on new points.
Classification Separation Surfaces for Vectors
Linear classifier, some degrees of freedom
Linear classifier, some degrees of freedom
Linear classifier, some degrees of freedom

Specially close to the border of the classifier
Linear classifier, some degrees of freedom
Linear classifier, some degrees of freedom

For each different technique, different results, different performance.
A few linear classifiers:
Linear Discriminant Analysis
A multivariate (＝ for vectors) generalization of the Gaussian density.

A very common density to characterize random vectors.
Reminder: Gaussian Multivariate Density

\[ \mathbf{x} \in \mathbb{R}^d \sim \mathcal{N}(\mu, \Sigma), \Sigma \text{ positive definite} \]

\[ \uparrow \]

Density of \( \mathbf{x} \) is
\[ \frac{1}{(2\pi)^{d/2}|\Sigma|} e^{(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)} \]
Reminder: Gaussian Multivariate Density
Linear Discriminant Analysis in a Nutshell

- Suppose we have points from two classes, 0 and 1
- Estimate Covariance $\Sigma_0, \Sigma_1$ and means $\mu_0$ and $\mu_1$ for each class
Linear Discriminant Analysis

- Decide that $x$ belongs to 1 if $p_1(x) > p_0(x)$ and vice versa.
- In practice this yields, after some computations, a decision of the form:
  - $x$ belongs to class 1 if
    \[
    (x - \mu_0)^T \Sigma_0^{-1} (x - \mu_0) + \ln |\Sigma_0| - (x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1) - \ln |\Sigma_1| > T
    \]
Linear Discriminant Analysis

- Decide that $x$ belongs to 1 if $p_1(x) > p_0(x)$ and vice versa.

- In practice this yields, after some computations, a decision of the form:
  - $x$ belongs to class 1 if
    \[
    (x - \mu_0)^T \Sigma_0^{-1} (x - \mu_0) + \ln |\Sigma_0| - (x - \mu_1)^T \Sigma_1^{-1} (x - \mu_1) - \ln |\Sigma_1| > T
    \]

- If one assumes that $\Sigma_0 = \Sigma_1 = \Sigma$, the decision becomes a simple dot-product:
  \[
  w^T x < C
  \]
  where
  \[
  w = \Sigma^{-1} (\mu_1 - \mu_0).
  \]
A few linear classifiers:
Logistic Regression
Regression does not work

- Consider the toy classification example:
  - Points $\mathbf{x}_j$ are taken randomly between -10 and 50.
  - The label
    \[
    y_j = \begin{cases} 
    0 & \text{if } x_j < \pi, \\
    1 & \text{if } x_j > \pi. 
    \end{cases}
    \]

- What happens if we feed this directly to regression?... **matlab demo**
How can we adapt regression? logistic map

- Logistic map:
  \[ g(z) = \frac{e^z}{e^z + 1} = \frac{1}{e^{-z} + 1} \]

○ for any \( z \), \( 0 \leq g(z) \leq 1 \)
How can we adapt regression? logistic map

Basic Idea

• Rather than find the best $c$ and $b$ such that

$$f(x_j) = c^T x_j + b \approx y_j \in \{0, 1\}$$

• logistic regression considers instead the best $c$ and $b$ such that

$$g \circ f(x_j) = \frac{1}{e^{-(c^T x_j + b)} + 1} \approx y_j \in \{0, 1\}.$$

• if for a new point $x$,

  ○ $g \circ f(x) > 1/2$, guess that the class is 1
  ○ $g \circ f(x) < 1/2$, guess that the class is 0
Probabilistic Interpretation of Logistic Regression

• Suppose there is a probability density \( p(X, Y) \) on couples \((x, y) \in \mathbb{R}^d \times \{0, 1\}\).

• Suppose for now that we know \( p \).

• The ratio

\[
 r(x) = \frac{p(Y = 1|X = x)}{p(Y = 0|X = x)}
\]

is called the odds-ratio of a given point \( x \).

• Obviously,
  
  o if \( r(x) > 1 \), then it is more likely that \( y = 1 \) than \( y = 0 \).
  o if \( r(x) < 1 \), then one is tempted to guess that \( y = 0 \) than \( y = 1 \).
Probabilistic Interpretation of Logistic Regression

• In other words...

\[
\log \frac{p(Y = 1|X = x)}{p(Y = 0|X = x)}, \quad \begin{cases} > 0 \text{ then } y = 1 \text{ is the likely answer} \\ < 0 \text{ then } y = 0 \text{ is the likely answer} \end{cases}
\]

• Logistic regression assumes that the log-odds ratio follows a linear relationship

\[
\log \frac{p(Y = 1|X = x)}{p(Y = 0|X = x)} \approx c^T x + b
\]

• This implies that the decision surface is linear.

Note that Logistic Regression assumes a model only for the log-odds ratio, not for the whole probability \( p \)
Probabilistic Interpretation of Logistic Regression

- Since \( p(Y = 0|X = x) = 1 - p(Y = 1|X = x) \), we hence have

\[
\log \frac{p(Y = 1|X = x)}{1 - p(Y = 1|X = x)} = c^T x + b
\]

- which in turn implies

\[
p(Y = 1|X = x) = \frac{1}{e^{-(c^T x + b)} + 1} = g(c^T x + b).
\]

Predictor variables contribute **linearly** to the increase/decrease of the probability that \( y = 1 \).
Estimation of $c$ and $b$ through Maximum Likelihood

- Flip coin, setting $p(y = 1) = p$ and $p(y = 0) = 1 - p$ for binary random variable $y$,
  - Likelihood of a draw $y$ knowing that probability is $p$,
    $$p^y(1 - p)^{1-y}$$

- In the context of **logistic regression**, $p$ depends on $c$, $b$ and $x_j$ for each point,
  $$\mathcal{L}(c, b) = \prod_{j=1}^{N} g(c^T x_j + b)^{y_j} (1 - g(c^T x_j + b))^{1-y_j}$$
Estimation of $c$ and $b$ through Maximum Likelihood

• Using again the log transformation,

$$\log L(c, b) = \sum_{j=1}^{N} y_j \log g(c^T x_j + b) + (1 - y_j) \log g(1 - (c^T x_j + b)).$$

• Maximizing this log-likelihood is equivalent to

$$\max_{c,b} \log L(c, b) \Leftrightarrow \max_{c,b} \sum_{j=1}^{N} y_j(c^T x_j + b) - \log(1 + e^{c^T x_j + b}).$$

• No closed form solution for this unfortunately... need efficient optimization.

• For datasets of reasonable size, Newton method for instance.
Estimation of $c$ and $b$ through Maximum Likelihood

Compare...

...with

\[ \frac{1}{\exp(-8.9576x + (-28.1031)) + 1} \]

\[ \frac{1}{2} (1 + \text{sign}(x - \pi)) \]
A few linear classifiers:
Perceptron Rule
Estimation of c and b through iterative updates

- Iterative algorithm that considers each point successively.
- Here we consider $S = \{-1, 1\}$
- Start from any arbitrary estimate $\omega = [\hat{b}]$.
- Loop over $j$ until $\omega$ does not change for a while...
  - Consider a point $[\frac{1}{x_j}]$ and his label $y_j$.
  - Do $u_j = \text{sign}(\omega^T [\frac{1}{x_j}])$ and $y_j$ match?
  - If not, set $\omega \leftarrow \omega + \rho(y_j - u_j) [\frac{1}{x_j}]$.
- Not much more to add, better see in practice.
A few linear classifiers: Support Vector Machine
A criterion to select a linear classifier: the margin?
A criterion to select a linear classifier: the margin?
A criterion to select a linear classifier: the margin?
A criterion to select a linear classifier: the margin?
A criterion to select a linear classifier: the margin?
Largest Margin Linear Classifier ?
Support Vectors with Large Margin
In equations

• We assume (for the moment) that the data are **linearly separable**, i.e., that there exists \((w, b) \in \mathbb{R}^d \times \mathbb{R}\) such that:

\[
\begin{align*}
\mathbf{w}^T \mathbf{x}_i + b &> 0 & \text{if } y_i = 1, \\
\mathbf{w}^T \mathbf{x}_i + b &< 0 & \text{if } y_i = -1.
\end{align*}
\]

• Next, we give a formula to compute the margin as a function of \(w\).

• Obviously, for any \(t \in \mathbb{R}\),

\[H_{w,b} = H_{tw,tb}\]

• Thus \(w\) and \(b\) are defined up to a multiplicative constant.

• We need to take care of this in the definition of the margin.
How to find the largest separating hyperplane?

For the linear classifier \( f(x) = w^T x + b \),
consider the interstice defined by the hyperplanes:

- \( f(x) = w^T x + b = +1 \)
- \( f(x) = w^T x + b = -1 \)

Consider \( x_1 \) and \( x_2 \) such that \( x_2 - x_1 \) is parallel to \( w \).
The margin is $2/\|\mathbf{w}\|$:

- Margin $= 2/\|\mathbf{w}\|$: the points $\mathbf{x}_1$ and $\mathbf{x}_2$ satisfy:

$$
\begin{align*}
\mathbf{w}^T \mathbf{x}_1 + b &= 0, \\
\mathbf{w}^T \mathbf{x}_2 + b &= 1.
\end{align*}
$$

- By subtracting we get $\mathbf{w}^T (\mathbf{x}_2 - \mathbf{x}_1) = 1$, and therefore:

$$
\gamma \overset{\text{def}}{=} 2\|\mathbf{x}_2 - \mathbf{x}_1\| = \frac{2}{\|\mathbf{w}\|}.
$$

where $\gamma$ is by definition the **margin**.
All training points should be on the appropriate side

- For positive examples \((y_i = 1)\) this means:
  \[
  \mathbf{w}^T \mathbf{x}_i + b \geq 1
  \]

- For negative examples \((y_i = -1)\) this means:
  \[
  \mathbf{w}^T \mathbf{x}_i + b \leq -1
  \]

- in both cases:
  \[
  \forall i = 1, \ldots, n, \quad y_i (\mathbf{w}^T \mathbf{x}_i + b) \geq 1
  \]
Finding the optimal hyperplane

- Find \((w, b)\) which minimize:

\[
\|w\|^2
\]

under the constraints:

\[
\forall i = 1, \ldots, n, \quad y_i (w^T x_i + b) - 1 \geq 0.
\]

This is a classical quadratic program on \(\mathbb{R}^{d+1}\)

**linear constraints - quadratic objective**
Another interpretation: Convex Hulls?

go back to 2 sets of points that are linearly separable
Another interpretation: Convex Hulls

Linearly separable = convex hulls do not intersect
Another interpretation: Convex Hulls

Find two closest points, one in each convex hull
Another interpretation: Convex Hulls

The SVM = bisection of that segment
Another interpretation: Convex Hulls

support vectors = extreme points of the faces on which the two points lie
The non-linearly separable case for SVM’s

(when convex hulls intersect)
What happens when the data is not linearly separable?
What happens when the data is not linearly separable?
What happens when the data is not linearly separable?
What happens when the data is not linearly separable?
Soft-margin SVM

- Find a trade-off between **large margin** and **few errors**.

- Mathematically:

  \[
  \min_f \left\{ \frac{1}{\text{margin}(f)} + C \times \text{errors}(f) \right\}
  \]

- \(C\) is a parameter
• The **margin** of a labeled point \((x, y)\) is

\[
\text{margin}(x, y) = y (w^T x + b)
\]

• The **error** is
  - 0 if \(\text{margin}(x, y) > 1\),
  - \(1 - \text{margin}(x, y)\) otherwise.

• The soft margin SVM solves:

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

• \(c(u, y) = \max\{0, 1 - yu\}\) is known as the **hinge loss**.

• \(c(w^T x_i + b, y_i)\) associates a mistake cost to the decision \(w, b\) for example \(x_i\).