Foundation of Intelligent Systems, Part I

Regression 2

mcuturi@i.kyoto-u.ac.jp
Some Words on the Survey

What is your main goal in taking this class?
Please check one or two boxes.

- I know nothing about machine learning, so I just need an introduction
- I know a few machine learning algorithms, but I would like to have a better theoretical understanding
- I know a few machine learning algorithms, but I would like to learn about more advanced ones
- I would like to understand how to use machine learning algorithms for a particular application (for instance, vision, bioinformatics etc.)

Not enough answers to say anything meaningful!

- Try again: survey.
Last Week

**Regression**: highlight a functional relationship between a *predicted variable* and *predictors*
Regression: highlight a functional relationship between a predicted variable and predictors

find a function $f$ such that

$\forall (x, y)$ that can appear, $f(x) \approx y$
Regression: highlight a functional relationship between a predicted variable and predictors to find an accurate function $f$ such that

$$\forall (x, y) \text{ that can appear }, f(x) \approx y$$

use a data set & the least-squares criterion:

$$\min_{f \in F} \frac{1}{N} \sum_{j=1}^{N} (y_j - f(x_j))^2$$
Last Week

**Regression**: highlight a functional relationship between a **predicted variable** and **predictors**

- when regressing a **real number** vs a **real number**:
  
  ![Scatter plot of Rent vs. Surface](image)

  - Least-Squares Criterion \( L(b, a_1, \cdots, a_p) \) to fit **lines**, polynomials.
  - results in solving a linear system.

  \[
  \frac{\partial^2 \text{order}(b, a_1, \cdots, a_p)}{\partial a_p} = \text{linear in } (b, a_1, \cdots, a_p)
  \]

  - When setting \( \frac{\partial L}{\partial a_p} = 0 \) we get \( p + 1 \) **linear** equations for \( p + 1 \) variables.
Regression: highlight a functional relationship between a predicted variable and predictors

- when regressing a real number vs \(d\) real numbers (vector in \(\mathbb{R}^d\)),
  - find best fit \(\alpha \in \mathbb{R}^n\) such that \((\alpha^T x + \alpha_0) \approx y\).
  - Add to \(d \times N\) data matrix, a row of 1’s to get the predictors \(X\).
  - The row \(Y\) of predicted values
  - The Least-Squares criterion also applies:
    \[
    L(\alpha) = \|Y - \alpha^T X\|^2 = \left(\alpha^T XX^T \alpha - 2YX^T \alpha + \|Y\|^2\right).
    \]
    \[
    \nabla_\alpha L = 0 \quad \Rightarrow \quad \alpha^* = (XX^T)^{-1} XY^T
    \]
- This works if \(XX^T \in \mathbb{R}^{d\times d}\) is invertible.
Last Week

\[ (X^*X') \setminus (X^*Y') \]

\[
\text{ans} = \begin{align*}
-0.049332605603095 & \times \text{age} \\
0.163122792160298 & \times \text{surface} \\
-0.004411580036614 & \times \text{distance} \\
2.731204399433800 & + 27.300 \text{ JPY}
\end{align*}
\]
Today

- A statistical / probabilistic perspective on LS-regression
- A few words on polynomials in higher dimensions
- A geometric perspective
- Variable co-linearity and Overfitting problem
- Some solutions: advanced regression techniques
  - Subset selection
  - Ridge Regression
  - Lasso
A (very few) words on the statistical/probabilistic interpretation of LS
• Assume that the values of $y$ are stochastically linked to observations $x$ as

$$y - (\alpha^T x + \beta) \sim \mathcal{N}(0, \sigma).$$

• This difference is a random variable called $\varepsilon$ and is called a residue.
The Statistical Perspective on Regression

• This can be rewritten as,

\[ y = (\alpha^T x + \beta) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma), \]

• We assume that the difference between \( y \) and \( (\alpha^T x + b) \) behaves like a Gaussian (normally distributed) random variable.

**Goal as a statistician:** Estimate \( \alpha \) and \( \beta \) given observations.
Identically Independently Distributed (i.i.d) Observations

- Statistical hypothesis: assume that the parameters are $\alpha = a, \beta = b$
Identically Independently Distributed (i.i.d) Observations

- Statistical hypothesis: assume that the parameters are $\alpha = a, \beta = b$
- In such a case, what would be the probability of each observation $(x_j, y_j)$?
Identically Independently Distributed (i.i.d) Observations

- Statistical hypothesis: **assuming that the parameters are** $\alpha = a, \beta = b$, what would be the **probability** of each observation?

  - For each couple $(x_j, y_j), j = 1, \cdots, N$,

  $$P(x_j, y_j \mid \alpha = a, \beta = b) = \frac{1}{\sqrt{2\pi\sigma}} \exp \left(-\frac{\|y_j - (a^T x_j + b)\|^2}{2\sigma^2}\right)$$
Identically Independently Distributed (i.i.d) Observations

- Statistical hypothesis: assuming that the parameters are $\alpha = a$, $\beta = b$, what would be the probability of each observation?:
  - For each couple $(x_j, y_j), j = 1, \cdots, N$,
    \[
P(x_j, y_j \mid \alpha = a, \beta = b) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left( -\frac{\|y_j - (a^T x_j + b)\|^2}{2\sigma^2} \right)
    \]
  - Since each measurement $(x_j, y_j)$ has been independently sampled,
    \[
P \left( \{(x_j, y_j)\}_{j=1,\cdots,N} \mid \alpha = a, \beta = b \right) = \prod_{j=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp \left( -\frac{\|y_j - (a^T x_j + b)\|^2}{2\sigma^2} \right)
    \]
Identically Independently Distributed (i.i.d) Observations

- Statistical hypothesis: assuming that the parameters are $\alpha = a, \beta = b$, what would be the probability of each observation?:
  - For each couple $(x_j, y_j), j = 1, \cdots, N$,
    \[
P(x_j, y_j | \alpha = a, \beta = b) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left( -\frac{\|y_j - (a^T x_j + b)\|^2}{2\sigma^2} \right)
    \]
  - Since each measurement $(x_j, y_j)$ has been independently sampled,
    \[
P \left( \{(x_j, y_j)\}_{j=1}^{N} | \alpha = a, \beta = b \right) = \prod_{j=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp \left( -\frac{\|y_j - (a^T x_j + b)\|^2}{2\sigma^2} \right)
    \]
  - A.K.A likelihood of the dataset $\{(x_j, y_j)_{j=1,\cdots,N}\}$ as a function of $a$ and $b$,
    \[
    \mathcal{L}_{\{(x_j,y_j)\}}(a,b) = \prod_{j=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp \left( -\frac{\|y_j - (a^T x_j + b)\|^2}{2\sigma^2} \right)
    \]
Hence, for $a, b$, the **likelihood** function on the dataset $\{(x_j, y_j)_{j=1, \ldots, N}\}$...

$$
\mathcal{L}(a, b) = \prod_{j=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp \left( -\frac{\|y_j - (a^T x_j + b)^2}{2\sigma^2} \right)
$$
Maximum Likelihood Estimation (MLE) of Parameters

Hence, for $a, b$, the **likelihood** function on the dataset $\{(x_j, y_j)_{j=1, \ldots, N}\}$...

$$L(a, b) = \prod_{j=1}^{N} \frac{1}{\sqrt{2\pi\sigma}} \exp \left( -\frac{\|y_j - (a^T x_j + b)\|^2}{2\sigma^2} \right)$$

Why not use the **likelihood** to **guess** $(a, b)$ given data?
Hence, for $a, b$, the **likelihood** function on the dataset $\{(x_j, y_j)_{j=1,\ldots,N}\}$...

$$
\mathcal{L}(a, b) = \prod_{j=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left( - \frac{\|y_j - (a^T x_j + b)\|^2}{2\sigma^2} \right)
$$

...the **MLE** approach selects the values of $(a, b)$ which **maximize** $\mathcal{L}(a, b)$
Maximum Likelihood Estimation (MLE) of Parameters

Hence, for $a, b$, the **likelihood** function on the dataset $\{(x_j, y_j)_{j=1,\ldots,N}\}$...

\[
\mathcal{L}(a, b) = \prod_{j=1}^{N} \frac{1}{\sqrt{2\pi\sigma}} \exp\left( -\frac{\|y_j - (a^T x_j + b)\|^2}{2\sigma^2} \right)
\]

...the **MLE** approach selects the values of $(a, b)$ which **maximize** $\mathcal{L}(a, b)$

- **Since** $\max_{(a,b)} \mathcal{L}(a, b) \Leftrightarrow \max_{(a,b)} \log \mathcal{L}(a, b)$

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

- **Hence** $\max_{(a,b)} \mathcal{L}(a, b) \Leftrightarrow \min_{(a,b)} \sum_{j=1}^{N} \|y_j - (a^T x_j + b)\|^2$...
Statistical Approach to Linear Regression

- Properties of the MLE estimator: convergence of $\|\alpha - a\|$?
- Confidence intervals for coefficients,
- Tests procedures to assess if model “fits” the data,

![Residues Histogram Relative Frequency](image)

- Bayesian approaches: instead of looking for one optimal fit $(a, b)$ juggle with a whole density on $(a, b)$ to make decisions
- etc.
A few words on polynomials in higher dimensions
A few words on polynomials in higher dimensions

- For $d$ variables, that is for points $\mathbf{x} \in \mathbb{R}^d$,
  - the space of polynomials on these variables up to degree $p$ is generated by
    \[
    \{ \mathbf{x}^\mathbf{u} \mid \mathbf{u} \in \mathbb{N}^d, \mathbf{u} = (u_1, \cdots, u_d), \sum_{i=1}^{d} u_i \leq p \}
    \]
    where the monomial $\mathbf{x}^\mathbf{u}$ is defined as $x_1^{u_1} x_2^{u_2} \cdots x_d^{u_d}$
  - Recurrence for dimension of that space: $\dim_{p+1} = \dim_p + \binom{p+1}{d+p}$
- For $d = 20$ and $p = 5$, $1 + 20 + 210 + 1540 + 8855 + 42504 > 50.000$

Problem with polynomial interpolation in high-dimensions is the explosion of relevant variables (one for each monomial)
Geometric Perspective
Recall the problem:

\[
X = \begin{bmatrix}
1 & 1 & \cdots & 1 \\
\vdots & \vdots & \ddots & \vdots \\
x_1 & x_2 & \cdots & x_N \\
\vdots & \vdots & \ddots & \vdots 
\end{bmatrix} \in \mathbb{R}^{d+1 \times N}
\]

and

\[
Y = [y_1 \quad \cdots \quad y_N] \in \mathbb{R}^N.
\]

We look for \(\alpha\) such that \(\alpha^T X \approx Y\).
Back to Basics

• If we transpose this expression we get $X^T\alpha \approx Y^T$,

$$
\begin{bmatrix}
1 & x_{1,1} & \cdots & x_{d,1} \\
1 & x_{1,2} & \cdots & x_{d,2} \\
\vdots & \vdots & \ddots & \vdots \\
1 & x_{1,k} & \cdots & x_{d,k} \\
\vdots & \vdots & \ddots & \vdots \\
1 & x_{1,N} & \cdots & x_{d,N}
\end{bmatrix}
\times
\begin{bmatrix}
\alpha_0 \\
\vdots \\
\alpha_d
\end{bmatrix}
=
\begin{bmatrix}
y_1 \\
\vdots \\
y_2 \\
\vdots \\
y_M
\end{bmatrix}
$$

• Using the notation $Y = Y^T$, $X = X^T$ and $X_k$ for the $(k+1)^{th}$ column of $X$,

$$
\sum_{k=0}^{d} \alpha_k X_k \approx Y
$$

• Note how the $X_k$ corresponds to all values taken by the $k^{th}$ variable.

• Problem: approximate/reconstruct Reconstructing $Y \in \mathbb{R}^N$ using $X_0, X_1, \cdots, X_d \in \mathbb{R}^N$?
Reconstructing $\mathbf{Y} \in \mathbb{R}^N$ using $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d$ vectors of $\mathbb{R}^N$.

- Our ability to approximate $\mathbf{Y}$ depends implicitly on the space spanned by $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d$.

Consider the observed vector in $\mathbb{R}^N$ of predicted values.
Reconstructing $\mathbf{Y} \in \mathbb{R}^N$ using $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d$ vectors of $\mathbb{R}^N$.

- Our ability to approximate $\mathbf{Y}$ depends implicitly on the space spanned by $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d$

Plot the first regressor $\mathbf{X}_0$...
Reconstructing $\mathbf{Y} \in \mathbb{R}^N$ using $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d$ vectors of $\mathbb{R}^N$.

- Our ability to approximate $\mathbf{Y}$ depends implicitly on the space spanned by $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d$.

Assume the next regressor $\mathbf{X}_1$ is colinear to $\mathbf{X}_0$...
Reconstructing $\mathbf{Y} \in \mathbb{R}^N$ using $\mathbf{X}_0, \mathbf{X}_1, \ldots, \mathbf{X}_d$ vectors of $\mathbb{R}^N$.

- Our ability to approximate $\mathbf{Y}$ depends implicitly on the space spanned by $\mathbf{X}_0, \mathbf{X}_1, \ldots, \mathbf{X}_d$ and so is $\mathbf{X}_2$...
Reconstructing \( Y \in \mathbb{R}^N \) using \( X_0, X_1, \ldots, X_d \) vectors of \( \mathbb{R}^N \).

- Our ability to approximate \( Y \) depends implicitly on the space spanned by \( X_0, X_1, \ldots, X_d \).

Very little choices to approximate \( Y \)...
Reconstructing $Y \in \mathbb{R}^N$ using $X_0, X_1, \cdots, X_d$ vectors of $\mathbb{R}^N$.

- Our ability to approximate $Y$ depends implicitly on the space spanned by $X_0, X_1, \cdots, X_d$

Suppose $X_2$ is actually not colinear to $X_0$. 
Reconstructing $\mathbf{Y} \in \mathbb{R}^N$ using $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d$ vectors of $\mathbb{R}^N$.

- Our ability to approximate $\mathbf{Y}$ depends implicitly on the space spanned by $\mathbf{X}_0, \mathbf{X}_1, \cdots, \mathbf{X}_d$

This opens new ways to reconstruct $\mathbf{Y}$.
Linear System

Reconstructing $Y \in \mathbb{R}^N$ using $X_0, X_1, \ldots, X_d$ vectors of $\mathbb{R}^N$.

- Our ability to approximate $Y$ depends implicitly on the space spanned by $X_0, X_1, \ldots, X_d$.

When $X_0, X_1, X_2$ are linearly independent,
Reconstructing $Y \in \mathbb{R}^N$ using $X_0, X_1, \cdots, X_d$ vectors of $\mathbb{R}^N$.

- Our ability to approximate $Y$ depends implicitly on the space spanned by $X_0, X_1, \cdots, X_d$.

$Y$ is in their span since the space is of dimension 3.
Reconstructing $Y \in \mathbb{R}^N$ using $X_0, X_1, \cdots, X_d$ vectors of $\mathbb{R}^N$.

- Our ability to approximate $Y$ depends implicitly on the space spanned by $X_0, X_1, \cdots, X_d$.

The dimension of that space is $\text{Rank}(X)$, the rank of $X$.

$\text{Rank}(X) \leq \min(d + 1, N)$. 
Three cases depending on \textbf{Rank X} and $d, N$

1. \textbf{Rank X} < $N$. \textit{$d + 1$ column vectors do not span} $\mathbb{R}^N$
   - For arbitrary $Y$, there is \textbf{no solution} to $\alpha^T X = Y$

2. \textbf{Rank X} = $N$ and $d + 1 > N$, \textbf{too many variables span the whole of} $\mathbb{R}^N$
   - \textbf{infinite} number of solutions to $\alpha^T X = Y$.

3. \textbf{Rank X} = $N$ and $d + 1 = N$, \textbf{# variables = # observations}
   - Exact and unique solution: $\alpha = X^{-1}Y$ we have $\alpha^T X = Y$

In most applications, $d + 1 \neq N$ so we are either in case 1 or 2
Case 1: Rank $X < N$

- **no solution** to $\alpha^T X = Y$ (equivalently $X\alpha = Y$) in general case.
- What about the **orthogonal projection** of $Y$ on the image of $X$.

\[\hat{Y} = \underset{u \in \text{span } \{X_0, X_1, \ldots, X_d\}}{\text{argmin}} \|Y - u\|\]

- Namely the point $\hat{Y}$ such that

$\text{span } \{X_0, X_1, \ldots, X_d\}$
Case 1: Rank $X < N$

Lemma 1. $\{X_0, X_1, \cdots, X_d\}$ is a l.i. family $\iff X^TX$ is invertible
Case 1: Rank $X < N$

- Computing the projection $\hat{\omega}$ of a point $\omega$ on a subspace $V$ is well understood.
- In particular, if $(X_0, X_1, \cdots, X_d)$ is a basis of $\text{span}\{X_0, X_1, \cdots, X_d\}$

  (that is $\{X_0, X_1, \cdots, X_d\}$ is a linearly independent family)

  ... then $(X^TX)$ is invertible and ...

  $$\hat{Y} = X(X^TX)^{-1}X^TY$$

- This gives us the $\alpha$ vector of weights we are looking for:

  $$\hat{Y} = X \underbrace{(X^TX)^{-1}X^T}_{\hat{\alpha}}Y = X\hat{\alpha} \approx Y \text{ or } \hat{\alpha}^TX = Y$$

- What can go wrong?
Case 1: Rank $X < N$

- If $X^TX$ is invertible,
  \[ \hat{Y} = X(X^TX)^{-1}X^T Y \]

- If $X^TX$ is not invertible... we have a problem.

- If $X^TX$’s condition number
  \[ \frac{\lambda_{\text{max}}(X^TX)}{\lambda_{\text{min}}(X^TX)} \]
  is very large, a small change in $Y$ can cause dramatic changes in $\alpha$.

- In this case the linear system is said to be **badly conditioned**...

- Using the formula
  \[ \hat{Y} = X(X^TX)^{-1}X^T Y \]
  might return garbage as can be seen in the following Matlab example.
Case 2: Rank $X = N$ and $d + 1 > N$

high-dimensional low-sample setting

- Ill-posed inverse problem, the set

$$\{ \alpha \in \mathbb{R}^d \mid X\alpha = Y \}$$

is a whole vector space. We need to choose one from many admissible points.

- When does this happen?
  - High-dimensional low-sample case (DNA chips, multimedia etc.)

- How to solve for this?
  - Use something called regularization.
A practical perspective:
Colinearity and Overfitting
A Few High-dimensions Low sample settings

- DNA chips are very long vectors of measurements, one for each gene

- Task: regress a health-related variable against gene expression levels

A Few High-dimensions Low sample settings

- Emails represented as bag-of-words email \( j = \begin{bmatrix}
  \vdots \\
  \text{please} = 2 \\
  \vdots \\
  \text{send} = 1 \\
  \vdots \\
  \text{money} = 2 \\
  \vdots \\
  \text{assignment} = 0 \\
  \vdots 
\end{bmatrix} \in \mathbb{R}^d \)

- Task: regress probability that this is an email against bag-of-words

Image: http://clg.wlv.ac.uk/resources/junk-emails/index.php
Correlated Variables

- Suppose you run a real-estate company.

- For each apartment you have compiled a few hundred predictor variables, e.g.
  - distances to conv. store, pharmacy, supermarket, parking lot, etc.
  - distances to all main locations in Kansai
  - socio-economic variables of the neighborhood
  - characteristics of the apartment

- Some are obviously correlated (correlated = “almost” colinear)
  - distance to Post Office / distance to Post ATM

- In that case, we may have some problems (Matlab example)

Source: http://realestate.yahoo.co.jp/
Overfitting

- Given \( d \) variables (including constant variable), consider the least squares criterion

\[
L_d (\alpha_1, \cdots, \alpha_d) = \sum_{i=1}^{j} \left\| y_j - \sum_{i=1}^{d} \alpha_i x_{i,j} \right\|^2
\]

- Add any variable vector \( x_{d+1,j}, j = 1, \cdots, N \), and define

\[
L_{d+1}(\alpha_1, \cdots, \alpha_d, \alpha_{d+1}) = \sum_{i=1}^{j} \left\| y_j - \sum_{i=1}^{d} \alpha_i x_{i,j} - \alpha_{d+1} x_{d+1,j} \right\|^2
\]
Overfitting

- Given $d$ variables (including constant variable), consider the least squares criterion

$$L_d(\alpha_1, \cdots, \alpha_d) = \sum_{i=1}^{j} \left\| y_j - \sum_{i=1}^{d} \alpha_i x_{i,j} \right\|^2$$

- Add any variable vector $x_{d+1,j}$, $j = 1, \cdots, N$, and define

$$L_{d+1}(\alpha_1, \cdots, \alpha_d, \alpha_{d+1}) = \sum_{i=1}^{j} \left\| y_j - \sum_{i=1}^{d} \alpha_i x_{i,j} - \alpha_{d+1} x_{d+1,j} \right\|^2$$

THEN $\min_{\alpha \in \mathbb{R}^{d+1}} L_{d+1}(\alpha) \leq \min_{\alpha \in \mathbb{R}^{d}} L_d(\alpha)$
Overfitting

- Given \( d \) variables (including constant variable), consider the least squares criterion

\[
L_d (\alpha_1, \cdots, \alpha_d) = \sum_{i=1}^{j} \left\| y_j - \sum_{i=1}^{d} \alpha_i x_{i,j} \right\|^2
\]

- Add any variable vector \( \mathbf{x}_{d+1,j}, j = 1, \cdots, N \), and define

\[
L_{d+1} (\alpha_1, \cdots, \alpha_d, \alpha_{d+1}) = \sum_{i=1}^{j} \left\| y_j - \sum_{i=1}^{d} \alpha_i x_{i,j} - \alpha_{d+1} x_{d+1,j} \right\|^2
\]

Then \( \min_{\alpha \in \mathbb{R}^{d+1}} L_{d+1}(\alpha) \leq \min_{\alpha \in \mathbb{R}^{d}} L_d(\alpha) \)

why? \( L_d (\alpha_1, \cdots, \alpha_d) = L_{d+1} (\alpha_1, \cdots, \alpha_d, 0) \)
Overfitting

- Given $d$ variables (including constant variable), consider the least squares criterion

$$L_d (\alpha_1, \cdots, \alpha_d) = \sum_{i=1}^{j} \left\| y_j - \sum_{i=1}^{d} \alpha_i x_{i,j} \right\|^2$$

- Add any variable vector $x_{d+1,j}, j = 1, \cdots, N$, and define

$$L_{d+1} (\alpha_1, \cdots, \alpha_d, \alpha_{d+1}) = \sum_{i=1}^{j} \left\| y_j - \sum_{i=1}^{d} \alpha_i x_{i,j} - \alpha_{d+1} x_{d+1,j} \right\|^2$$

Then $\min_{\alpha \in \mathbb{R}^{d+1}} L_{d+1}(\alpha) \leq \min_{\alpha \in \mathbb{R}^d} L_d(\alpha)$

why? $L_d (\alpha_1, \cdots, \alpha_d) = L_{d+1} (\alpha_1, \cdots, \alpha_d, 0)$

**Residual-sum-of-squares** goes down... but is it relevant to add variables?
Occam’s razor formalization of overfitting

Minimizing least-squares (RSS) is not clever enough. We need another idea to avoid overfitting.

- **Occam’s razor**: *lex parsimoniae*

- **law of parsimony**: principle that recommends selecting the hypothesis that makes the fewest assumptions.

  *one should always opt for an explanation in terms of the fewest possible causes, factors, or variables.*

Advanced Regression Techniques
Quick Reminder on Vector Norms

- For a vector $a \in \mathbb{R}^d$, the Euclidian norm is the quantity

$$\|a\|_2 = \sqrt{\sum_{i=1}^{d} a_i^2}.$$ 

- More generally, the q-norm is for $q > 0$,

$$\|a\|_q = \left( \sum_{i=1}^{d} |a_i|^q \right)^{\frac{1}{q}}.$$ 

- In particular for $q = 1$,

$$\|a\|_1 = \sum_{i=1}^{d} |a_i|$$

- In the limit $q \to \infty$ and $q \to 0$,

$$\|a\|_\infty = \max_{i=1,\ldots,d} |a_i|.$$ 

$$\|a\|_0 = \# \{i \mid a_i \neq 0\}.$$
Tikhonov Regularization ’43 - Ridge Regression ’62

- Tikhonov’s motivation: solve **ill-posed inverse problems** by **regularization**
- If \( \min_\alpha L(\alpha) \) is achieved on many points... consider
  \[
  \min_\alpha L(\alpha) + \lambda \|\alpha\|^2_2
  \]
- We can show that this leads to selecting
  \[
  \hat{\alpha} = (X^T X + \lambda I_{d+1})^{-1} X Y
  \]
- The condition number has changed to
  \[
  \frac{\lambda_{\max}(X^T X) + \lambda}{\lambda_{\min}(X^T X) + \lambda}.
  \]
Subset selection : Exhaustive Search

• Following Ockham’s razor, ideally we would like to know for any value $p$

$$\min_{\alpha, \|\alpha\|_0=p} L(\alpha)$$

• → select the best vector $\alpha$ which only gives weights to $p$ variables.

• → Find the best combination of $p$ variables.

Practical Implementation

• For $p \leq n$, $\binom{n}{p}$ possible combinations of $p$ variables.

• Brute force approach: generate $\binom{n}{p}$ regression problems and select the one that achieves the best RSS.

Impossible in practice with moderately large $n$ and $p...\binom{30}{5} = 150.000$
Subset selection: Forward Search

Since the exact search is **intractable in practice**, consider the **forward** heuristic

- **In Forward search:**
  - define $I_1 = \{0\}$.
  - given a set $I_k \subset \{0, \cdots, d\}$ of $k$ variables, **what is the most informative variable one could add?**
    - Compute for each variable $i$ in $\{0, \cdots, d\} \setminus I_k$
      
      $$t_i = \min_{(\alpha_k)_{k \in I_k}, \alpha} \sum_{j=1}^{N} \left\| y_j - \left( \sum_{k \in I_k} \alpha_k x_{k,j} + \alpha x_{i,j} \right) \right\|^2$$
    - Set $I_{k+1} = I_k \cup \{i^*\}$ for any $i^*$ such that $i^* = \min t_i$.
    - $k = k + 1$ until desired number of variables
Subset selection: Backward Search

... or the **backward** heuristic

**In Backward search:**

- define $I_d = \{0, 1, \cdots, n\}$.
- given a set $I_k \subset \{0, \cdots, d\}$ of $k$ variables, **what is the least informative variable** one could **remove**?
  - Compute for each variable $i$ in $I_k$

$$t_i = \min_{(\alpha_k)_{k \in I_k \setminus \{i\}}} \sum_{j=1}^{N} \left\| y_j - \left( \sum_{k \in I_k \setminus \{i\}} \alpha_k x_{k,j} \right) \right\|^2$$

- Set $I_{k-1} = I_k \setminus \{i^*\}$ for any $i^*$ such that $i^* = \max t_i$.
- $k = k - 1$ until desired number of variables
Subset selection: LASSO

Naive Least-squares

$$\min_{\alpha} L(\alpha)$$

Best fit with \( p \) variables (Occam!)

$$\min_{\alpha, \|\alpha\|_0=p} L(\alpha)$$

Tikhonov regularized Least-squares

$$\min_{\alpha} L(\alpha) + \lambda \|\alpha\|_2^2$$

LASSO (least absolute shrinkage and selection operator)

$$\min_{\alpha} L(\alpha) + \lambda \|\alpha\|_1$$