Probabilistic Models for Supervised Learning

Logistic Regression & Conditional Random Field

Song Liu
song@sg.cs.titech.ac.jp
JSPS PD, Sugyiama Lab.
Tokyo Institute of Technology

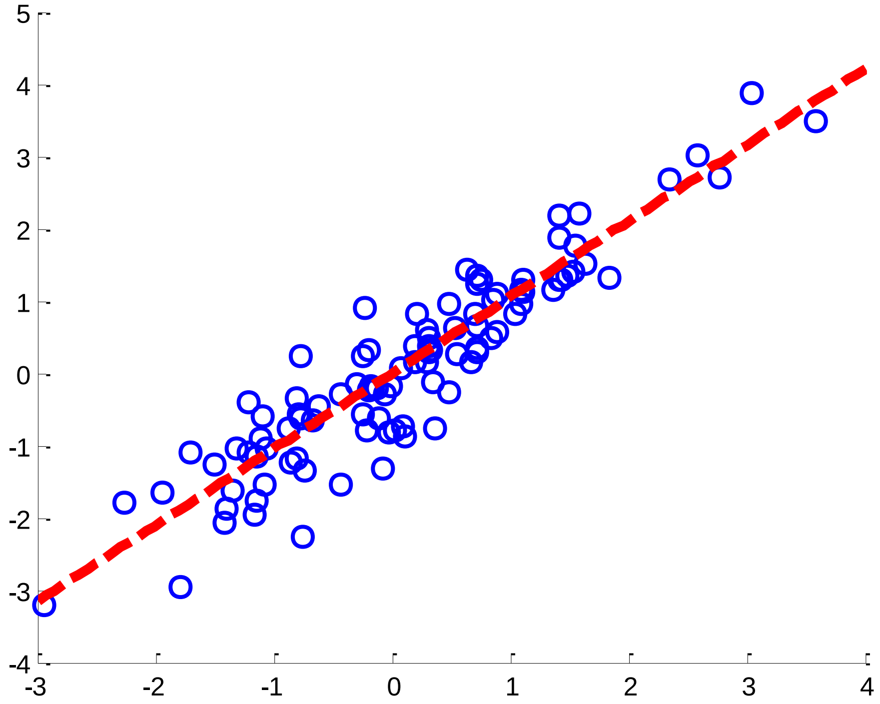
Essentially, all models are wrong,
but some are useful.

George E. P. Box

Notations

- $\boldsymbol{x} \in R^p$ p-dimensional covariates, predictive features
- $\boldsymbol{X} \in R^{p \times n}$, $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n]$ data/design matrix
- $y \in R$ response variable for regression
    - or $y \in \{0,1\}$ response variable for classification
- $\boldsymbol{\beta} \in R^p$ regression coefficient
- $\epsilon \sim N(0, \sigma^2)$ i.i.d. noise

The Good Old Least Squares…



Objective:

$$\min_{\boldsymbol{\beta}} || \boldsymbol{y} - \boldsymbol{X}^\top \boldsymbol{\beta} ||^2$$

Data generated by

$$y = \boldsymbol{x}^\top \boldsymbol{\beta} + \epsilon$$

Solution:

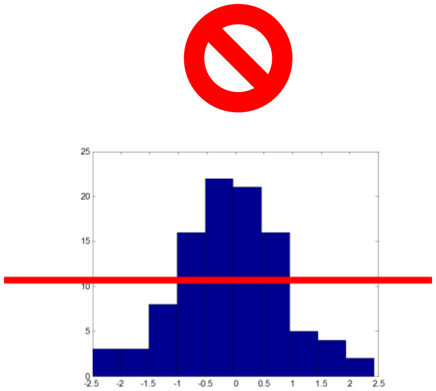
$$\boldsymbol{\beta} = (\boldsymbol{X}\boldsymbol{X}^\top)^{-1} \boldsymbol{X}\boldsymbol{y}^\top$$

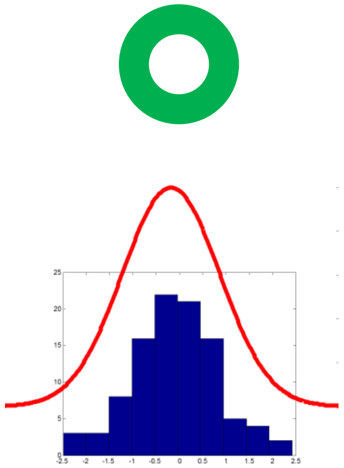
What is the *probabilistic model* behind?

Before introducing probabilistic *models*, let’s first look at probabilistic *algorithms*.

The Maximum Likelihood Estimator (MLE)

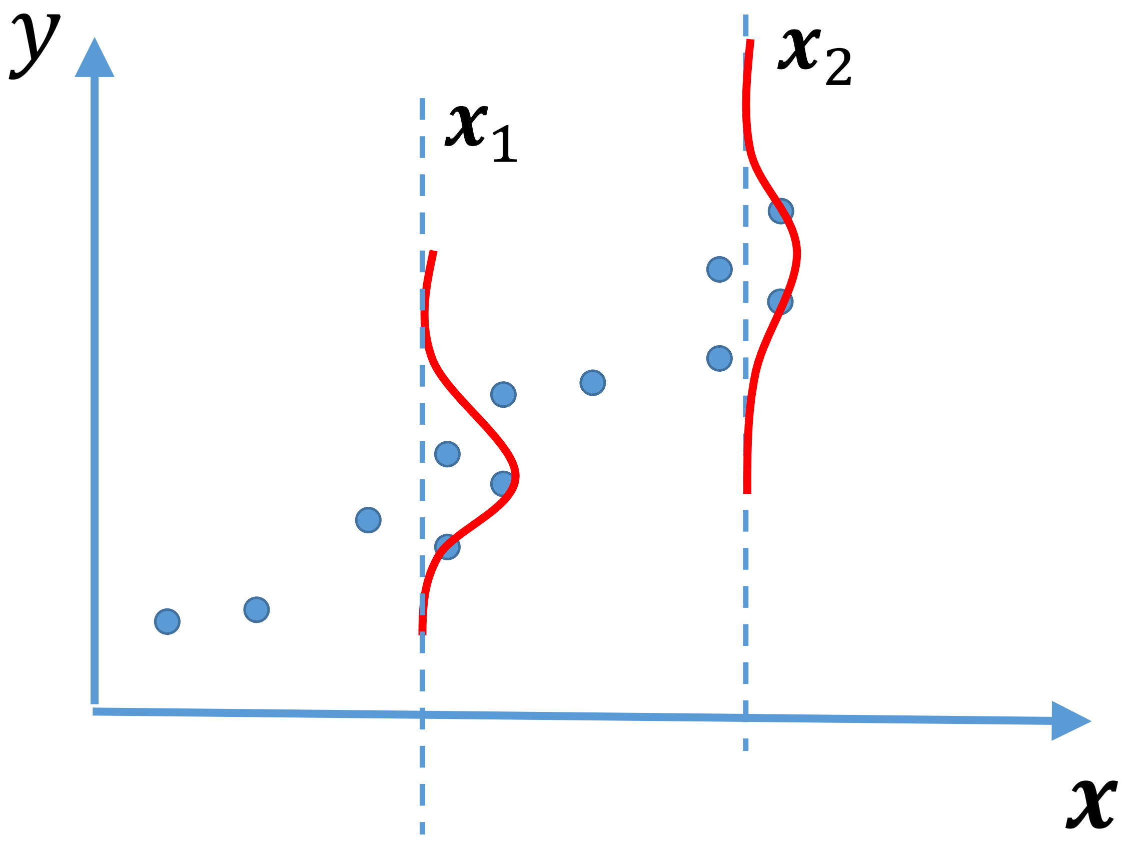
- Given samples $\{\boldsymbol{z}_i\}_{i=1}^{n} iid \sim p(\boldsymbol{z})$,
- and a model $p(\boldsymbol{z}|\boldsymbol{\beta})$,
- MLE finds estimates of $\boldsymbol{\beta}$.





- $\max_{\boldsymbol{\beta}} \frac{1}{n} \sum_i \log p(\boldsymbol{z}_i|\boldsymbol{\beta})$
  - Intuitively, maximizing the agreement between the model and observations.

We are interested in $p(y|\boldsymbol{x})$ in supervised learning…



$\boldsymbol{x}$ is location, building years, number of rooms … , y is the house price.

The Maximum **Conditional** Likelihood Estimator

- Given samples $\{(\boldsymbol{x_i}, y_i)\}_{i=1}^{n} iid \sim p(\boldsymbol{x}, y)$,
- and a model $p(y|\boldsymbol{x}; \boldsymbol{\beta})$,
  - Note that $\boldsymbol{x}$ is behind the bar.
  - p is defined on y alone.

- $\max_{\boldsymbol{\beta}} \frac{1}{n} \sum_i \log p(y_i|\boldsymbol{x}_i; \boldsymbol{\beta})$
  - Or you may think x is just another parameter.

The man behind MLE

Sir **Ronald Aylmer Fisher** (17 February 1890 – 29 July 1962) was an English statistician, evolutionary biologist, geneticist, and eugenicist.

Now, let’s talk about **models**.

What if We Combine MLE and Gaussian Density Model?

- $p(y|\boldsymbol{x};\boldsymbol{\beta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{||y-\boldsymbol{x}^\top\boldsymbol{\beta}||^2}{2\sigma^2}}$, σ is known (don't care).
- MLE becomes:
- $\max_{\boldsymbol{\beta}} \frac{1}{n}\sum_i \log p(y_i|\boldsymbol{x}_i;\boldsymbol{\beta})$

constant

$$= \frac{1}{n}\sum_i \log(e^{-\frac{||y_i-\boldsymbol{x}_i^\top\boldsymbol{\beta}||^2}{2\sigma^2}}) - \boxed{\log\sqrt{2\pi\sigma^2}}$$

$$= \frac{1}{n}\sum_i -\frac{||y_i-\boldsymbol{x}_i^\top\boldsymbol{\beta}||^2}{2\sigma^2} - \log\sqrt{2\pi\sigma^2}$$

Least-squares is MLE + Gaussian density model.

Gauss-Markov Model

Recall, least-squares assumes data are generated by

$$y = \boldsymbol{x}^{\top}\boldsymbol{\beta} + \epsilon, \text{ where } \epsilon \text{ i.i.d.} \sim N(0, \sigma^2)$$

Least-squares can be fit into a probabilistic Alg. + Model.

What if data are not generated via the above model?

What if data are discrete?

We need a more general paradigm.

Exponential Family (log-linear model)

- A wide range of probabilistic models are similar in definition:
- $p(z;\boldsymbol{\theta}) = \boxed{p_0(z)} \exp(\boldsymbol{\theta}^\top \boxed{\boldsymbol{f}(z)} - \log N(\boldsymbol{\theta}))$

Base measure   Sufficient stats

- $N(\boldsymbol{\theta}) = \int_Z p_0(z) \exp\left(\boldsymbol{\theta}^\top \boldsymbol{f}(z)\right) dy$

Normalization function ensures $\int p(z;\theta) dz \ = 1$

- Examples: Normal, Gamma, Poisson Distribution…
- Such model is sometimes called "log-linear model".

Exponential Family (conditional)

- Conditional densities can also be expressed via Exponential Family model.
- Just use $z = (y, \boldsymbol{x})$, and **normalize w.r.t. $\boldsymbol{y}$.**
- $p(y|\boldsymbol{x}, \boldsymbol{\theta}) = \exp(\boldsymbol{\theta}^\top \boldsymbol{f}(y, \boldsymbol{x}) - \log N(\boldsymbol{\theta}; \boldsymbol{x}))$

$$N(\boldsymbol{\theta}; \boldsymbol{x}) = \int_Y \exp\left(\boldsymbol{\theta}^\top \boldsymbol{f}(y, \boldsymbol{x})\right) dy$$

Normalization function ensures $\int_Y p(y|\boldsymbol{x}, \boldsymbol{\theta}) dy = 1$.

Probability Model: 1) Positive, 2) Normalized

Exponential Family (conditional)

- How does Normal distribution fit into this paradigm? (Suppose $y, x \in R$)
- QUIZ: what are $f, \boldsymbol{\theta}$ in this case?
- $p(y|\boldsymbol{x}; \boldsymbol{\beta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{||y - \boldsymbol{x}^\top \boldsymbol{\beta}||^2}{2\sigma^2}}$
- $p(z; \boldsymbol{\theta}) = p_0(z) \exp(\boldsymbol{\theta}^\top \boldsymbol{f}(z) - g(\boldsymbol{\theta}))$
- $\boldsymbol{f}(y, x) = \begin{bmatrix} y^2 \\ yx \\ x^2 \end{bmatrix}, \boldsymbol{\theta} = \begin{bmatrix} 1/2\sigma^2 \\ \beta/\sigma^2 \\ \beta^2/2\sigma^2 \end{bmatrix}.$

Why Exponential Family?

- Models from Exponential Family are highly expressive (as we will show).
- The resulting optimization problem is **convex**
  - **No local optimal!**
  - **Simple gradient descent will do!**

We first extend the such idea to classification problems.

Binary Classification

- Now think about classification problems: $y \in \{0,1\}$.
- y is now binary. We need to know
  - $p(y=1|\boldsymbol{x})$ or $1-p(y=1|\boldsymbol{x})$

- Prob. distribution for *binary random variables*?
- Bernoulli Distribution!
  - Bernoulli distribution is "flipping a coin".
  - Imagining "training a smart coin".
  - Multi-class classification?
  - Train a smart dice…

Bernoulli Distribution

- Its prob. mass function is given by:
- $P(z; m) = m^z (1-m)^{1-z}, z = \{0,1\}$

- Bernoulli distribution is also a member of **Exponential Family.**
    - Let $\theta = \log \frac{m}{1-m}$
- $P(z; \theta) = \frac{\exp(z \cdot \theta)}{\boxed{1+\exp(\theta)}}$

Normalization term

- or $P(z; \theta) = \exp(z \cdot \theta \; - \boxed{\log(1 + \exp \theta)})$

Logistic Regression

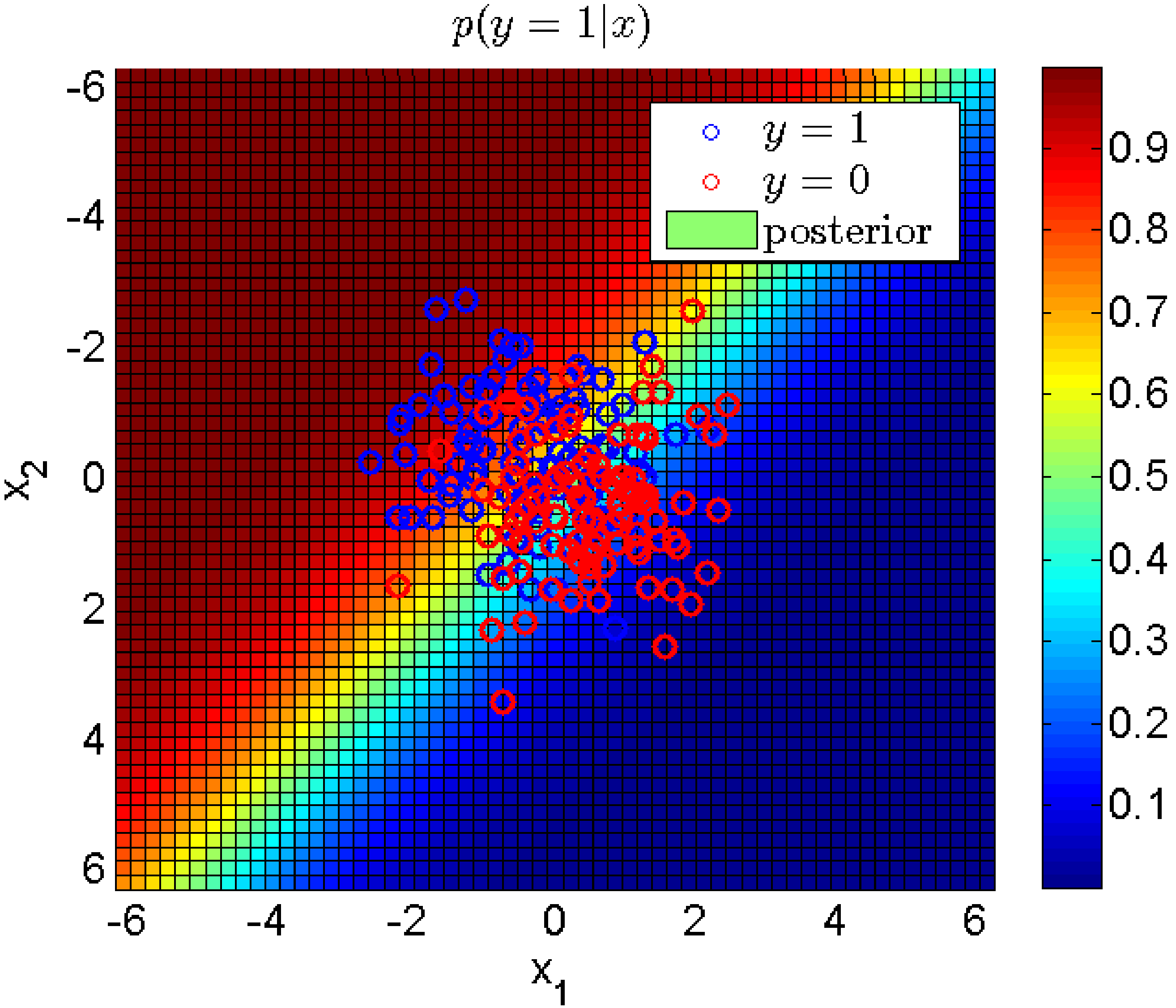
- We need an model of class posterior, i.e.
- $p(y|\boldsymbol{x},\boldsymbol{\theta}),$ where $y = \{0,1\}$.
- $P(y|\boldsymbol{x},\boldsymbol{\theta}) = \frac{\exp(y \cdot \boldsymbol{\theta}^\top \boldsymbol{x})}{1+\exp(\boldsymbol{\theta}^\top \boldsymbol{x})}$

Hint: by substituting $z = (\boldsymbol{x}, y)$

- Again, use MLE algorithm
- $\max\limits_{\boldsymbol{\theta}} \frac{1}{n} \sum_i \log p(y_i|\boldsymbol{x}_i,\boldsymbol{\theta})$
$= \frac{1}{n} \sum_i (y_i \cdot \boldsymbol{\theta}^\top \boldsymbol{x}_i - \log(1 + \exp(\boldsymbol{\theta}^\top \boldsymbol{x}_i)))$

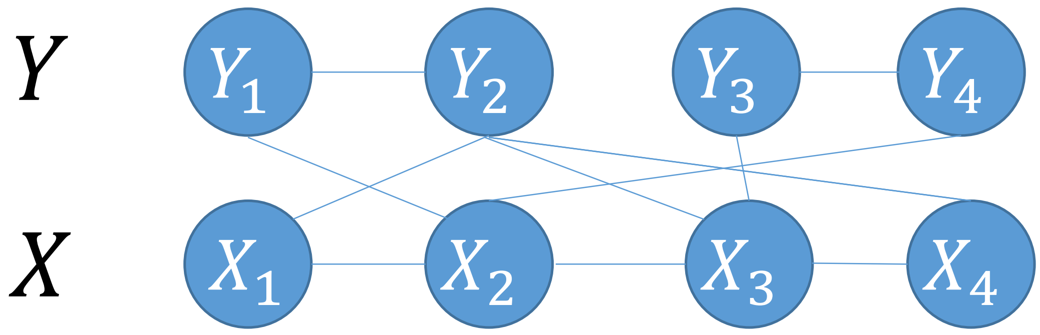
The resulting algorithm is called Logistic-Regression.

Logistic Regression, Example



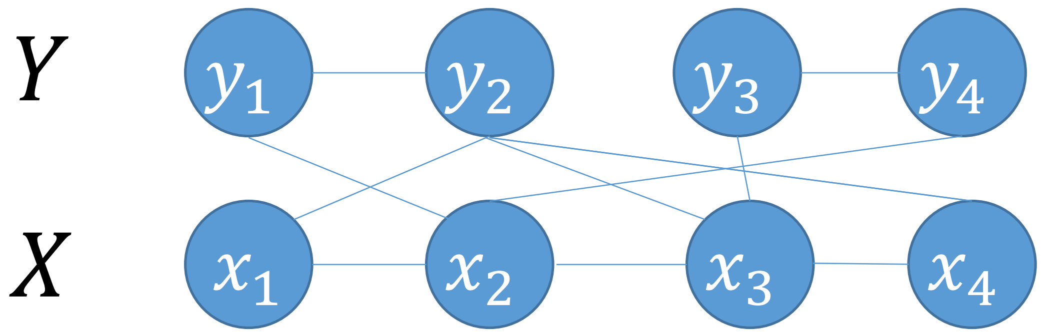
Predicting Label Vectors

- What if the label is not a simple binary variable?
- For example:
  - $\boldsymbol{x} \in R^p, \boldsymbol{y} \in \{0,1\}^p$
    - $\boldsymbol{x}$ and $\boldsymbol{y}$ are both vector now.
- Imagine $\boldsymbol{y}$ and $\boldsymbol{x}$ have some **structures**, say, a **graph**.



- This setting will bring some interesting applications.

Markov Random Field (MRF)



- The edges in the graph indicates **conditional dependence.**
- For an undirected graph, $G = < V, E >$, Z is a set of random variables indexed by V.
- $(A, B) \notin E$, if $A \perp B \mid_{Z \setminus \{A,B\}}$
- Links can be roughly understood as "interactions" between random variables.

Gene Prediction

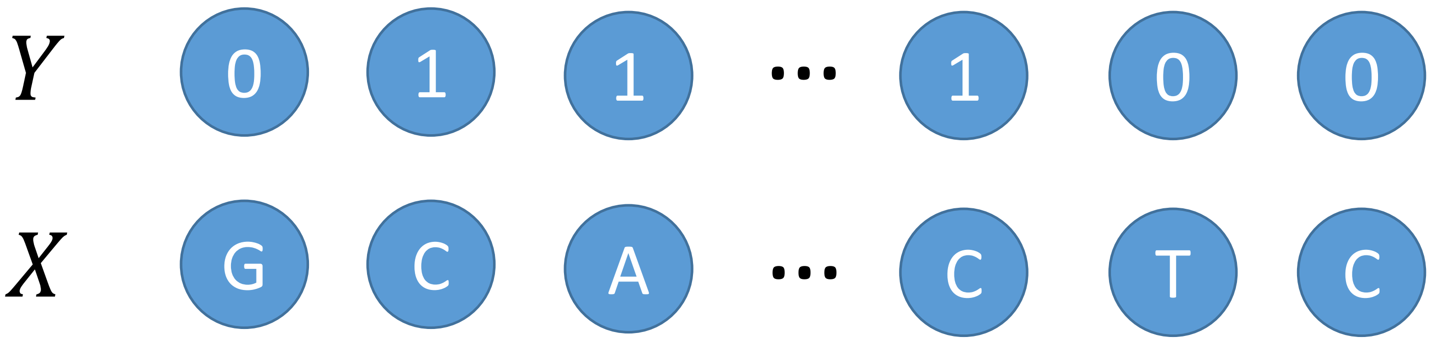
- A case study: Genetic Coding.
- DNA sequence is a sequence of nucleic acids. “DNA markup” is a string with repeating “A”,”T”,”C” and “G”, used to represent DNA sequence in text.
- DNA carries “blueprints” of proteins.
- Only some segments of DNA sequence contains “blueprints” for proteins, called **GENEs**.

Gene Prediction

```
>chromosome:GRCh37:13:32889011:32974405:1
TACCAAGCCCTGCGGAGCAAGGTACCTCACACTTCATGAGCGAGTTAAGATGGGTTTCAC
AATTTTTCAAGCAAGGAAACGGGCTCGGAGGTCTTGAACACCTGCTACCCAATAGCAGAA
CAGCTACTGGAACTAAAATCCTCTGATTTCAAATAACAGCCCCGCCCACTACCACTAAGT
GAAGTCATCCACAACCACACACCGACCACTCTAAGCTTTTGTAAGATCGGCTCGCTTTGG
GGAACAGGTCTTGAGAGAACATCCCTTTTAAGGTCAGAACAAAGGTATTTCATAGGTCCC
AGGTCGTGTCCCGAGGGCGCCCACCCAAACATGAGCTGGAGCAAAAAGAAAGGGATGGGG
GACTTGGAGTAGGCATAGGGGCGGCCCCTCCAAGCAGGGTGGCCTGGGACTCTTAAGGGT
CAGCGAGAAGAGAACACACACTCCAGCTCCCGCTTTATTCGGTCAGATACTGACGGTTGG
GATGCCTGACAAGGAATTTCCTTTCGCCACACTGAGAAATACCCGCAGCGGCCCACCCAG
GCCTGACTTCCGGGTGGTGCGTGTGCTGCGTGTCGCGTCACGGCGTCACGTGGCCAGCGC
GGGCTTGTGGCGCGAGCTTCTGAAACTAGGCGGCAGAGGCGGAGCCGCTGTGGCACTGCT
GCGCCTCTGCTGCGCCTCGGGTGTCTTTTGCGGCGGTGGGTCGCCGCCGGGAGAAGCGTG
AGGGGACAGATTTGTGACCGGCGCGGTTTTTGTCAGCTTACTCCGGCCAAAAAAGAACTG
CACCTCTGGAGCGGGTTAGTGGTGGTGGTAGTGGGTTGGGACGAGCGCGTCTTCCGCAGT
CCCAGTCCAGCGTGGCGGGGGAGCGCCTCACGCCCCGGGTCGCTGCCGCGGCTTCTTGCC
CTTTTGTCTCTGCCAACCCCCACCCATGCCTGAGAGAAAGGTCCTTGCCCGAAGGCAGAT
TTTCGCCAAGCAAATTCGAGCCCCGCCCCTTCCCTGGGTCTCCATTTCCCGCCTCCGGCC
CGGCCTTTGGGCTCCGCCTTCAGCTCAAGACTTAACTTCCCTCCCAGCTGTCCCAGATGA
CGCCATCTGAAATTTCTTGGAAACACGATCACTTTAACGGAATATTGCTGTTTTGGGGAA
GTGTTTTACAGCTGCTGGGCACGCTGTATTTGCCTTACTTAAGCCCCTGGTAATTGCTGT
ATTCCGAAGACATGCTGATGGGAATTACCAGGCGGCGTTGGTCTCTAACTGGAGCCCTCT
```

Gene Prediction

- Task: Labelling genes from DNA markups.



The **exact mapping rules** from X to Y have been unknown to scientists yet, and perhaps is **very complicated**.

- However, expert labelled X and Y pairs are available.
- We can learn a **probabilistic model**!

“Far better an **approximate answer** to the **right** question, which is often vague, than an **exact answer** to the **wrong** **question**, which can always be made precise.”

John Tukey

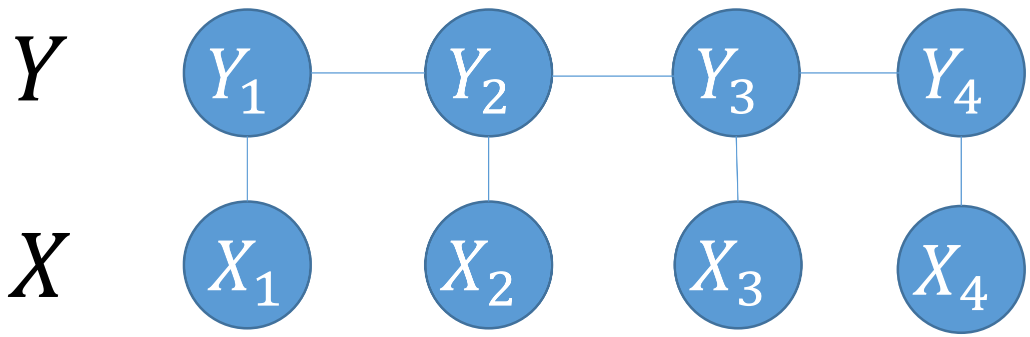
Another Example

- Part of Speech (POS) Labelling

Y	PRONOUN	VERB	PARTICLE	LOCATION
X	I	live	in	Tokyo.

- Labelling the *lexical properties* in a sentence
- Important for computer to extract key information
- For example, **named-entities.**
  - Locations, Person Names, Company Names…

Probabilistic Model for Sequences



- Suppose, Y is an underlying hidden variable.
  - e.g. Gene label (0: non-gene, 1: gene).
- X is an observed variable, generated from Y.
  - e.g. the DNA sequence, "ATGCG..."
- Given paired samples $(\boldsymbol{x}, \boldsymbol{y})$, we may learn a model:
  - $p(\boldsymbol{y}|\boldsymbol{x}; \beta)$.
- By using such model, given an observed $\boldsymbol{x}'$, we may infer a possible label $\boldsymbol{y}$'.

Conditional Random Field (CRF)

- In the previous example, X are not linked between each other, and Y are only linked as a **chain**.
- X and Y can have more complicated structures, depending on applications.
- Generally speaking, a probabilistic model $p(\boldsymbol{y}|\boldsymbol{x};\beta)$ defined on a Markov Random Field on $Z = X \cup Y$, is called **conditional random field**.
  - This model is very suitable for supervised learning.
  - “discriminative model”

Modelling CRF

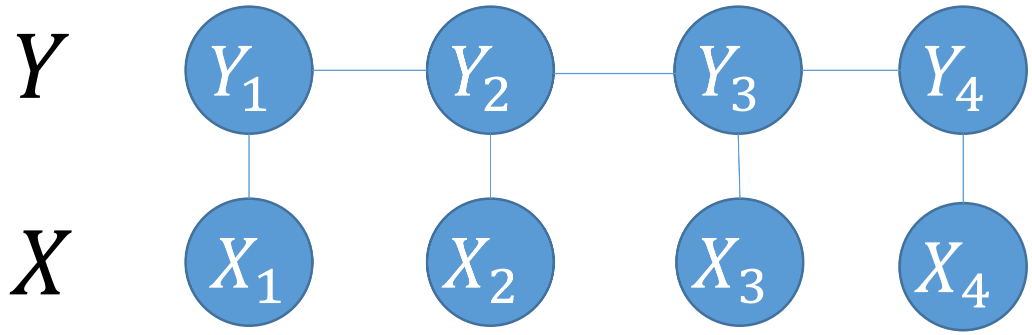
- Can we fit CRF into **Exponential Family**?
  - If so, learning CRF would be similar to the learning of earlier models, by using gradient descent.
- YES, we CAN!
  - MRF itself is a member of Exponential Family.
  - The log-linear model of MRF is sometimes called **Gibbs distribution**.

Modelling CRF

- The exponential family has the following form:
- $p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}) = \exp(\boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{y},\boldsymbol{x}) - \log N(\boldsymbol{\theta};\boldsymbol{x}))$
  - The question is, how to design $\boldsymbol{f}$?
  - $\boldsymbol{f}$ needs to capture **the intrinsic information** of x and y.
- Can't we define one feature jointly on $\boldsymbol{x}$ and $\boldsymbol{y}$?
  - Yes, we can! e.g. $\boldsymbol{f}: R^{p\times p} \to R$
- However, $\boldsymbol{y}$ and $\boldsymbol{x}$ are both p dimensional vectors.
  - Design such feature function may be hard.
  - Only a scalar output is not expressive enough.

Modelling CRF

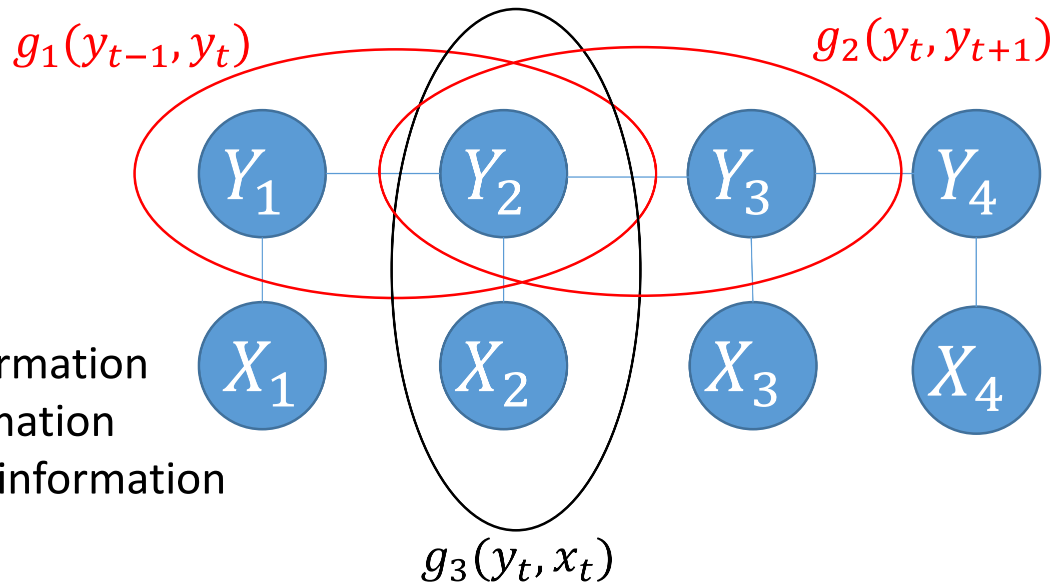
- For simplicity, we only consider chain shaped CRF:



- It is suggested that to use the following $\boldsymbol{f}$ for a chain-shaped CRF:

Modelling CRF

Extract sufficient statistics only on linked pairwise-random variables



g_1: backward information
g_2: forward information
g_3: observational information

- For example, whether the current label position is a named entity depends on
  - Whether the previous word is a Name Suffix(“Mr. or Mrs.”) ?
  - Whether the next word is a Company Suffix(“Inc.”)
  - Does the current word start with a capital letter (“Tokyo”)?

Modelling CRF

"weights of features"

- $$\boldsymbol{\theta} = \begin{bmatrix} 0 \\ \theta_2 \\ \theta_3 \\ \dots \\ \theta_1 \\ \theta_2 \\ \theta_3 \\ \dots \\ \theta_1 \\ 0 \\ \theta_3 \end{bmatrix}, \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}) = \begin{bmatrix} 0 \\ g_2(y_2, y_1) \\ g_3(x_1, y_1) \\ \dots \\ g_1(y_{i-1}, y_i) \\ g_2(y_i, y_{i+1}) \\ g_3(x_i, y_i) \\ \dots \\ g_1(y_{p-1}, y_p) \\ 0 \\ g_3(x_p, y_p) \end{bmatrix}$$

If the labelling is not **position specific**, we can **share parameters**.

Helps when sequences have different lengths!

- We may hand-craft as many feature as we like, and
- let the data speak for itself!

Modelling CRF

- How to choose g heavily depending on applications.
  - CRF provides great flexibility on choosing features!
- However, in the simplest case $g(z_1, z_2) = z_1 \cdot z_2$.

Learning CRF

- Like other supervised learning tasks, we want to learn parameter $\boldsymbol{\theta}$ in the probability model $p(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta})$.
- Using MLE, we have the following learning objective:
- $\max_{\boldsymbol{\theta}} \frac{1}{n}\sum_i \log p(y_i|\boldsymbol{x}_i,\theta)$

$$= \frac{1}{n}\sum_i \sum_t \begin{matrix} \theta_1 g_1\left(y_{t-1}^{(i)}, y_t^{(i)}\right) \\ +\theta_2 g_2\left(y_t^{(i)}, y_{t+1}^{(i)}\right) \\ +\theta_3 g_3(y_t^{(i)}, x_t^{(i)}) \end{matrix} - \log N\left(\theta_1,\theta_2,\theta_3,\boldsymbol{x}^{(i)}\right)$$

Sample index is (i)

Position index is t

Note that only 3 parameters need to be estimated.

However, what is N?

The Pain of Normalization

- $N(\theta_1, \theta_2, \theta_3, \boldsymbol{x_i}) =$
$$\sum_{y} \exp\left(\sum_i \sum_t \begin{matrix} \theta_1 g_1\left(y_{t-1}^{(i)}, y_t^{(i)}\right) \\ +\theta_2 g_2\left(y_t^{(i)}, y_{t+1}^{(i)}\right) \\ +\ \theta_3 g_3\left(y_t^{(i)}, x_t^{(i)}\right) \end{matrix} \right)$$
- N is the normalization term that guarantees the probability is summed up to one.
- An unfortunate thing is, the summation is over the entire domain of y.

The Pain of Normalization

- How large is the entire domain of $\boldsymbol{y}$?
- Imagine that $\boldsymbol{y}$ is a sequence of p binary digits, then $\boldsymbol{y} \in \{0,1\}^p$.
- There are 2^p possible configurations of $\boldsymbol{y}$.
- BTW, the number of atoms in universe is around 2^{256}.[1]
- Predict long sequences by using this model is not possible.

1. http://en.wikipedia.org/wiki/Observable_universe#Matter_content_.E2.80.94_number_of_atoms

The Pain of Normalization

- The solution to this problem is beyond the scope of this class.
- Please refer to the book Daphne & Friedman, 2009, Chapter 20.6 for details.

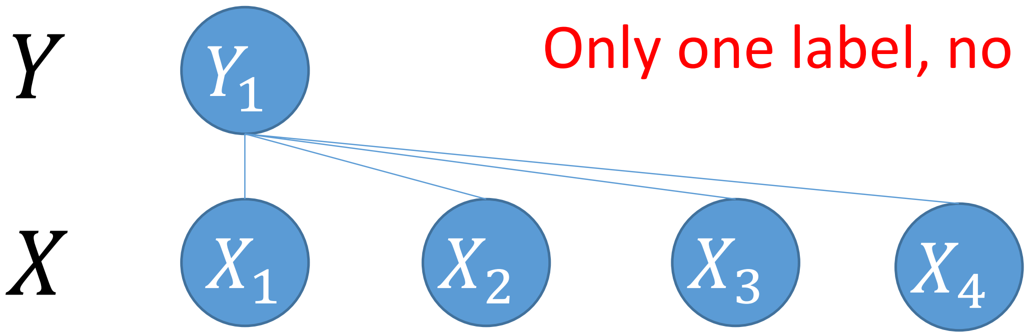
- D. Koller and N. Friedman (2009). **Probabilistic Graphical Models: Principles and Techniques**. edited by . MIT Press.

Logistic Regression, a Look Back

- Recall the Logistic regression use the model:

$$P(y|\boldsymbol{x}, \boldsymbol{\theta}) = \frac{\exp(y \cdot \boldsymbol{\theta}^\top \boldsymbol{x})}{1 + \exp(\boldsymbol{\theta}^\top \boldsymbol{x})}$$

- Logistic Regression is in fact, a very simple conditional random field, with
  - $g(x_t, y) = x_t \cdot y$



Only one label, no sequence!

Logistic Regression, a Look Back

$$P(y|\boldsymbol{x}, \boldsymbol{\theta}) = \frac{\exp(y \cdot \boldsymbol{\theta}^\top \boldsymbol{x})}{1 + \exp(\boldsymbol{\theta}^\top \boldsymbol{x})}$$

- Note, since the label of logistic regression only take two values, i.e.
  - $y \in \{0,1\}$
- Therefore, it only sums up over two summands, and is no problem in normalization.

Conclusion

- Probabilistic models for supervised learning tasks:
  - Gauss-Markov Model (regression)
  - Logistic Regression (classification)
  - Conditional Random Fields (sequence labelling)
- A unified framework
  - Maximize the conditional likelihood + Probabilistic Models from Exponential Family
    - Highly Expressive
    - Convex

Take-home Messages:

Least Squares

Logistic Regression

Conditional Random Field

are Maximal Likelihood Estimators of

Posterior Probability

$$p(Y|X)$$

Further Readings

- For label predictions using linear models, and their extensions:
- http://www.is.titech.ac.jp/~s-taiji/lecture/dataanalysis/L4.pdf
- 「データ解析」, by Prof. Suzuki, in Japanese.
- For introductions of Conditional Random Field
  - Lafferty et al., 2001,
    - Conditional random fields: Probabilistic models for segmenting and labeling sequence data
  - Daphne & Frideman, 2009
    - Chapter 20.3.2

Further Readings

- For fun reading, anecdotes in statistics.
  - David Salsburg, 2001
    - The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century

Homework

Write you opinion about the special lecture today.
Directly submit the printed report to the lecturer next week.