This lecture presents a brief overview of statistical models in machine learning, exponential family of distributions and the generalized linear model.

1 Statistical Models

In machine learning, data could appear in various forms, for example:

- Data → vectors, $x \in \mathbb{R}^p$ (continuous or discrete)
- → matrices, $x \in \mathbb{R}^{m \times n}$ (images)
- → trees (evolutionary phylogenetic trees)

The broad perspective from which we shall be viewing data is that:

Data $\rightarrow$ Outcome of a random experiment.

Let’s model the data as a random variable $x$, $x \in \mathcal{X}, \mathcal{X} \subset \mathbb{R}^p$ (continuous) OR $\mathcal{X} \subset \{0, 1\}^p$ (binary vector)

So, if we denote the probability distribution function of $x$ as $P(x)$, $P(x) \in \{ P(x; \theta) ; \theta \in \Theta \}$ (Statistical Model)

Examples of some common distributions:

1. Univariate Gaussian distribution:
   
   $P(x; \mu; \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}e^{\frac{1}{2\sigma^2}(x - \mu)^2}$

2. Multivariate Gaussian distribution:
   
   $P(X; \mu; \Sigma) = \frac{1}{(2\pi)^{d/2}\det(\Sigma)^{\frac{1}{2}}}e^{\frac{1}{2}(X - \mu)^T\Sigma^{-1}(X - \mu)}$

3. Bernoulli distribution (discrete random variables, $x \in \{0, 1\}$)
   
   $P(x; \theta) = \theta^x(1 - \theta)^{1-x}$
2 Exponential Family of Distributions

Let’s consider a general family of distribution in its standard form, which we shall call the Exponential Family of Distributions.

\[ P(x; \theta) = h(x) \exp(\eta(\theta)^T T(x) - A(\theta)) \]

where \( h \) is a function of \( x \), \( \eta \) is a function of the parameter \( \theta \), \( T \) is the sufficient statistics and \( A \) is the normalizing constant.

We can specify an exponential family via the tuple \((h, \eta, T, \Theta)\)

The log-normalizing constant is \( \exp(A(\theta)) \), where

\[ \exp(A(\theta)) = \int h(x) \exp(\eta(\theta)^T T(x)) dx, \theta \in \Theta, \eta: \Theta \to \mathbb{R}^k \]

Let us try to identify some common distributions as exponential family distributions:

1. Binomial distribution:

\[ P(x; \theta) = \binom{n}{x} \theta^x (1 - \theta)^{n-x} = \binom{n}{x} \exp(x \log \theta + (n - x) \log(1 - \theta)) = \binom{n}{x} \exp(x \log \frac{\theta}{1-\theta} + n \log(1 - \theta)) \]

where \( h(x) = \binom{n}{x} \), \( T(x) = x \), \( \eta(\theta) = \log \left( \frac{\theta}{1-\theta} \right) \), \( A(\theta) = n \log(1 - \theta) \)

2. Poisson distribution:

\[ P(x; \theta) = \frac{1}{x!} \theta^x e^{-\theta}, \theta > 0 \]

\[ = \frac{1}{x!} \exp(x \log \theta - \theta) \]

where \( h(x) = \frac{1}{x!}, T(x) = x, \eta(\theta) = \log \theta, A(\theta) = \theta \)

3. Gaussian distribution:

\[ P(x; \mu, \sigma^2) = \frac{1}{\sqrt{2 \pi \sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2}) \]

\[ = \exp(-\frac{x^2}{2\sigma^2} - \frac{\mu^2}{2\sigma^2} + \frac{\mu x}{\sigma^2} + \log \left( \frac{1}{\sqrt{2 \pi \sigma^2}} \right)) \]

\[ = \exp \left( \begin{pmatrix} x \end{pmatrix}^T \begin{pmatrix} -\frac{1}{\sigma^2} \\ \frac{\mu}{\sigma^2} \end{pmatrix} - \frac{\mu^2}{2\sigma^2} + \log \left( \frac{1}{\sqrt{2 \pi \sigma^2}} \right) \right) \]

where \( T(x) = \begin{pmatrix} x \end{pmatrix}^T, \eta(\theta) = \begin{pmatrix} -\frac{1}{\sigma^2} \\ \frac{\mu}{\sigma^2} \end{pmatrix}, A(\theta) = \frac{\mu^2}{2\sigma^2} - \log \left( \frac{1}{\sqrt{2 \pi \sigma^2}} \right) \)

2.1 Canonical Exponential Family

The exponential family of distributions can be more conveniently written in a compact form as
\[ P(x; \eta) = h(x) \exp(\eta^T T(x) - A(\eta)) \]

In other words, instead of taking the function of parameter, just take the parameter for representing the model.

For example, consider the Bernoulli distribution,
\[ P(x; \theta) = \exp(x \log \frac{\theta}{1 - \theta} - \log(1 - \theta)), \text{ canonical form is } \exp(\eta x - A(\eta)) \]

3 Prediction

Consider the pair \((X, Y)\), where \(X\) is our data and \(Y\) is the label on the corresponding data. \(X\) is also called the input / feature / covariate / dependent variable and \(Y\) is the output / response variable.

Now, consider the conditional probability distribution \(P(Y|X; \theta)\). If \(Y\) is discrete, then the task of prediction is called a classification problem and if \(Y\) is continuous, then it is called a regression problem.

Clearly \(P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}\), where \(Y\) is discrete, say \(Y \in \{C_1, C_2\}\)

Now, by Bayes Rule, we can expand it as:
\[ P(Y = C_1|X) = \frac{P(X|Y=C_1)P(Y=C_1)}{P(X|Y=C_1)P(Y=C_1) + P(X|Y=C_2)P(Y=C_2)} \]

This can also be written in a compact form:
\[ = \frac{\exp(a)}{1 + \exp(a)} \]

where \(a = \log \frac{P(X|Y=C_1)P(Y=C_1)}{P(X|Y=C_2)P(Y=C_2)} \) (log-odds)

Let’s consider the case when \(Y\) is continuous, \(Y \in \mathbb{R}^p\)

Say, \((X, Y)\) is jointly Gaussian, that is \((X, Y) \sim \mathcal{N}(\mu, \Sigma)\), where
\[
\mu = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \Sigma = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}
\]

As an exercise, find out \(P(Y|X)\), when it is known that \((X, Y)\) is jointly Gaussian. (Hint: should be a \(\mathcal{N}(?, ?)\) - find out the parameters of this normal distribution).

The above is the generative model for prediction where the joint distribution can be modeled as
\[ P(X, Y; \theta) = h(x) \exp(\eta^T T(x) - A(\theta)) \text{ [exponential family]} \]

and \(P(Y|X; \theta)\) can be computed using Bayes Rule.

Contrast this with the discriminative model of prediction where we are interested in \(P(Y|X) \in \{P(Y|X; \theta); \theta \in \Theta\}\)
4 Linear Models (Regression Model)

Let us study linear models of the form \( y = \theta^T x + \epsilon \), where \( y \in \mathbb{R}, \theta \in \mathbb{R}^p, \epsilon \in \mathcal{N}(0, \sigma^2) \) and \( x \in \mathbb{R}^p \).

So, \( P(Y|X; \theta) = \mathcal{N}(\theta^T x, \sigma^2) \)

Some examples of linear models are:

1. Logistic Regression Model:
   \( Y \in \{0, 1\} \) and \( P(Y = 1|X; \theta) = \frac{\exp(\theta^T x)}{1+\exp(\theta^T x)} \)
   which is in the general form as \( \frac{\exp(a)}{1+\exp(a)} = \frac{1}{1+\exp(-a)} = \sigma(a) \) (sigmoid function).

2. Generalized Linear Models (model conditional probability distribution as exponential families)
   \( P(Y|\theta) = h(Y) \exp(\eta(\theta)^T Y - A(\theta)) \)
   Given, \( \mathbb{E}[Y] = \mu(\theta) = g^{-1}(\beta^T X) \), where \( g \) is known as the link function. Therefore, \( \theta = \mu^{-1} g^{-1}(\beta^T x) \).
   So, \( P(Y|X, \beta) = h(Y) \exp(\mu^{-1} g^{-1}(\beta^T x) T(Y) - A(\beta)) \).

Now, set \( g = \mu^{-1} \), then the canonical form of the generalized linear model becomes:
   \( P(Y|X, \beta) = h(Y) \exp((\beta^T x) T(Y) - A(\beta)) \).

Exercise:

How can logistic regression be represented as a canonical generalized linear model?