COS513: FOUNDATIONS OF PROBABILISTIC MODELS
LECTURE 8: Linear Regression

Scribed by: Tian Long Wang

October 24, 2010

1 Probabilistic generative process

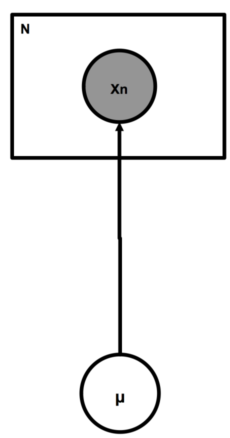

Figure 1: Generative Model

In probability and statistics, a generative model is a model for randomly generating observable data given some hidden variable parameters. It specifies a joint probability distribution over observed variables and hidden parameters. Figure 1 shows a graphical model representation of generating data points from a mean variable.

1. $\mu \sim N(\mu_0, \tau^2)$ - generate μ from a prior μ_0.

2. $X_n|\mu \sim N(\mu, \sigma^2)$ - generative process

We want to look at the posterior inference - $p(\mu|X_1, ..., X_n)$

We can estimate the hidden variable μ using maximum likelihood - $\hat{\mu} = arg \max_\mu logp(X_1, ..., X_n|\mu)$

2 Mixture Model

Mixture Model is a probabilistic model for density estimation using a mixture distribution. This is another example of widely used generative process.

1. $\mu_k \sim N(\mu_0, \tau^2)$ for k = 1, ... , K. Where k is the indexing of component, K is the total number of components.

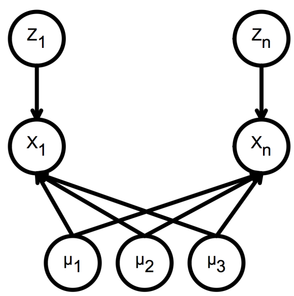

Figure 2: Mixture Model

2. For each datapoint:
   (a) Choose $Z_n \sim Discrete(\pi)$ where π represents a uniform distribution over $1, ..., k$.
   (b) Choose $X_n \sim \mathcal{N}(\mu_{Z_n}, \sigma^2)$ as shown in Figure 2.

   We are interested in $p(\mu_2|X_1, ..., X_n)$. However, we can see that all z_i, i $\in \{1, ..., n\}$ are dependent on each other. Therefore, we will need approximate inference to compute this.

3 Regression

In some models, we always observe and condition on certain aspects of the data. Our purpose is to maximize the conditional likelihood.

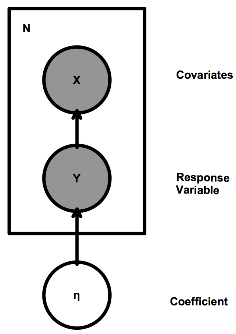

Figure 3: Regression

As shown in Figure 3, we have the following relationships:
The uncertainties on Y_n is modeled though a Gaussian distribution.

$$Y_n \sim \mathcal{N}(\eta^T x_n, \sigma^2)$$

The parameter estimator is the one that maximum the likelihood of parameter η.

$$\hat{\eta} = arg \max_{\eta} \sum_n log\ p(y_n|x_n, \eta)$$
$$\because p(\eta|X_{1:N}, Y_{1:N}) \propto p(\eta) \prod_n p(y_n|x_n, \eta)$$

N.B. When we condition on X_n, the model will be a discriminative model.

With X and Y representing different kind of data, we have different type of regression. For example:

$$X anything, Y continuous => linear\ regression$$
$$X anything, Y categorical => soft - max\ regression$$
$$X anything, Y binary => logistic\ regression$$

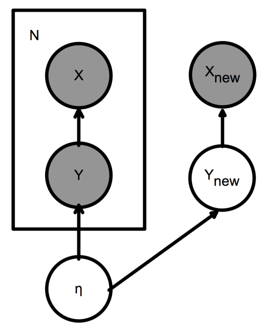

Figure 4: Regression Model with new variable to predict

We are interested in per-data prediction, this is illustrated in Figure 4.

The frequentist view of predicting y_{new} is $p(Y_{new}|X_{new}, \hat{\eta})$ where $\hat{\eta}$ is the parameter estimator using maximum likelihood.

The bayesian way of predicting y_{new} is the following (the conditional independencies can be obtained from the graphical model in Figure 4:

$$\begin{aligned} p(Y_{new}|X_{new}, \mathcal{D}) &= \int p(Y_{new}, \eta|X_{new}, \mathcal{D})\, d\eta \\ &= \int p(Y_{new}, |\eta, X_{new}, \mathcal{D}) p(\eta|X_{new}, \mathcal{D})\, d\eta \\ \because Y_{new} \perp\!\!\!\perp \mathcal{D}|\eta \;\; \& \;\; \eta \perp\!\!\!\perp X_{new}|\phi \therefore &= \int p(Y_{new}, |\eta, X_{new}) p(\eta|\mathcal{D})\, d\eta \end{aligned}$$

4 Ways of organizing models

In probabilistic modeling, there are several ways of organizing models:

1. Bayesian vs. Frequentist.
2. Discriminative vs. Generative.
   (a) Discriminative: conditioned on some variables
   (b) Generative: we fit a probability distribution to every part of the data, e.g. clustering, naive Bayesian classification.

3. Per-data point prediction vs. Data set density estimation.
4. Supervised vs. Unsupervised models.
   (a) Supervised: given $\{(x_i, y_i)\}_{i=1}^N$ in training, predict y given x in testing (e.g. classification).
   (b) Unsupervised: given data, we seek the structure of it. e.g. Clustering

However, all of these boundaries are soft. All of these models involves treat observations as random variables in a model. Solve our problem with a probabilistic computation about the model.

5 Linear Regression

In this section, we will talk about the basic idea of linear regression and then study how to fit a linear regression.

5.1 Overview

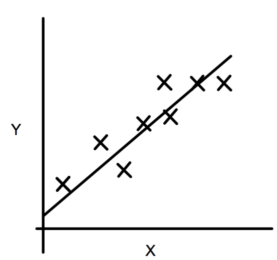

Figure 5: Linear regression. 'X's are data points and the dashed line is the output of fitting the linear regression.

The goal of Linear regression is to predict a real value response form a set of inputs (or covariates). See Figure 5 shows an example. Usually, we have multiple *covariates* $X_n =< X_{1,n}, X_{2,n}, \dots, X_{p,n} >$, where p is the number of covariates, n is number of covariates.

In linear regression, we fit a linear function of covariates

$$f(x) = \beta_0 + \sum_{i=1}^{p} \beta_i x_i = \beta_0 + \beta^T x.$$

Note that in general $\beta^T x = 0$ is a hyperplane.

Many candidate features can be used as the input x:

1. any raw numeric data;
2. any transformation, e.g. $x_2 = \log x_1$ and $x_3 = \sqrt{x_1}$;
3. basis expansions, e.g. $x_2 = x_1^2$ and $x_3 = x_1^3$;
4. indicator functions of qualitative inputs, e.g. 1[the subject has brown hair]; and
5. interactions between other covariates, e.g. $x_3 = x_1 x_2$.

5.2 Fitting a linear regression

Suppose we have a dataset $D = \{(x_n, y_n)\}_{n=1}^{N}$. In the simplest form of a linear regression, we assume $\beta_0 = 0$ and $p = 1$. So the function to be fitted is just

$$f(x) = \beta x.$$

To fit a linear regression in this simplified setting, we minimize the sum of the distances between fitted values and the truth. Thus, the objective function is

$$\text{RSS}(\beta) = \frac{1}{2} \sum_{n=1}^{N} (y_n - \beta x_n)^2,$$

Thus we can estimate β like this.

$$\hat{\beta} = arg \min_{\beta} \sum_{n=1}^{N} (y_n - \beta x_n)^2$$