CSE 788.04: Topics in Machine Learning **Lecture Date: April 9, 2012**

Lecture 5: Bayesian Linear Regression

Lecturer: Brian Kulis *Scribe: Jimmy Voss*

1 Linear Regression

1.1 Classical Approach

Given input pairs: $\begin{matrix} x_1, \ x_2, \ \cdots, \ x_n & x_i \in \mathbb{R}^n \\ y_1, \ y_2, \ \cdots, \ y_n & y_i \in \mathbb{R} \end{matrix}$

In linear regression, one models the predicted values for the y_is as linear combinations of their features:

$$\begin{aligned} x_i &= (x_i(1), \cdots, x_i(d))^T \\ \hat{y}_i &= w_0 + w_1 x_i(1) + \cdots + w_d x_i(d) \\ \hat{y}_i &= w_0 + w^T x \end{aligned}$$

To compress the notation further, this can be written without any w_0 term by mapping:

$$x_i \mapsto \begin{pmatrix} x_i \\ 1 \end{pmatrix} \qquad w_i \mapsto \begin{pmatrix} w_i \\ w_0 \end{pmatrix}$$

Then, we write $\hat{y}_i = w^T x_i$ for the linear regression equation. Often, the observed value for y_i is considered to be made up of 2 components:

$$y_i = w^T x_i + \epsilon \tag{1}$$

where ϵ is considered to be the error term. Error is assumed to be Gaussian noise: $\epsilon \sim N(0, \sigma^2)$. This allows for the following probability model of y_i:

$$\begin{aligned} P(y_i | w, X, \sigma^2) &= N(w^T x_i, \sigma^2) \\ P(y | w, X, \sigma^2) &= \prod_i P(y_i | w, x_i, \sigma^2) \end{aligned}$$

Classically, this likelihood function is maximized with respect to w. Defining

$$X = \begin{pmatrix} --- & x_1^T & --- \\ --- & x_2^T & --- \\ & \vdots & \\ --- & x_n^T & --- \end{pmatrix}$$

Notice that each x_i is a row vector in X rather than a column vector. To find the maximum likelihood estimate of w, one typically maximizes the log-likelihood:

$$\begin{aligned}\ln(P(y|w, x, \sigma^2) &= \sum_i \ln(N(y_i|w, x_i, \sigma^2)) \\ &= \sum_i \left[-\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - w^T x_i)^2 \right]\end{aligned}$$

Then the maximization problem is equivalent to finding:

$$w_{MLE} = \operatorname*{argmin}_w \frac{1}{2}\sum_{i=1}^{n}\left(y_i - x_i^T w\right)^2 \tag{2}$$

differentiating the right hand side with respect to w and setting the resulting equation equal to 0 in order to obtain critical points yields:

$$\begin{aligned}\sum_{i=1}^{n}\left(y_i - x_i^T w\right) x_i^T &= 0 \\ w &= (X^T X)^{-1} X^T y\end{aligned}$$

1.2 Treating Regression as an Optimization problem

Notice that solving for w_{MLE} in equation (2) can be viewed as an optimization problem. Treating (2) as an optimization problem can be useful since it can get rid of numerical issues such as non-invertible matrices. It also allows for a penalty to be placed upon using complicated weight vectors w. 2 such techniques follow:

1. In **Ridge Regression**, the following equation is minimized (w.r.t. w):

$$\frac{1}{2}\sum_{i=1}^{n}(y_i - w^T x_i)^2 + \frac{\lambda}{2}w^T w \tag{3}$$

Here a regularization term in the L_2 norm is used. The result is:

$$w = (\lambda I + X^T X)^{-1} X^T y \tag{4}$$

2. **Lasso** – The following equation is minimized (w.r.t. w):

$$\frac{1}{2}\sum_{i=1}^{n}(y_i - w^T x_i)^2 + \lambda\sum_{i=1}^{n}|w_i|$$

thus using an L_1 norm regularization term. This often results in solutions that are partially sparse (that is, many of the w_i terms are 0).

2 Bayesian Linear Regression

In the Bayesian case, one puts prior distributions on one or both of w and σ^2.

2.1 w is given a prior distribution

The conjugate prior for w is the normal distribution: $P(w) \sim N(\mu_0, S_0)$. To compute the posterior distribution:

$$\begin{aligned} P(w|y) &\propto P(y|w)P(w) \\ P(w|y) &\sim N(\mu, S) \\ S^{-1} &= S_0^{-1} + \frac{1}{\sigma^2}X^T X \\ \mu &= S\left(S_0^{-1}\mu_0 + \frac{1}{\sigma^2}X^T y\right) \end{aligned}$$

If one assumes that $S_0 = (1/\alpha)I$ and $\mu_0 = 0$, then

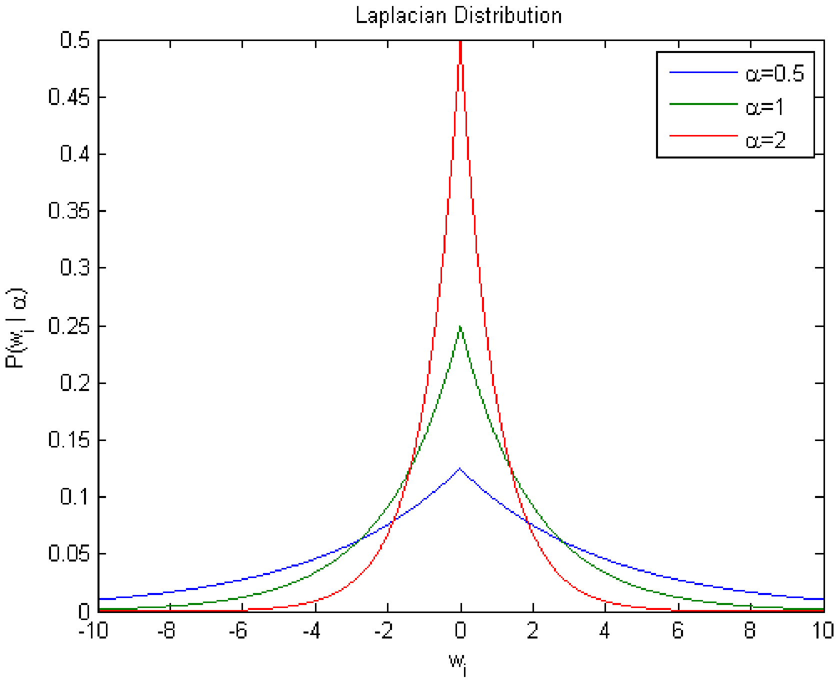
$$\begin{aligned} \mu &= \left(\alpha I + \frac{1}{\sigma^2}X^T X\right)^{-1}\left(\frac{1}{\sigma^2}X^T y\right) \\ &= \left(\alpha\sigma^2 I + X^T X\right)^{-1} X^T y \end{aligned}$$

which is equivalent to the weight vector equation found via Ridge Regression in equation (4). Looking at the maximum a posteriori probability (MAP) estimate yields the same result.

Alternatively, using the Laplace Distribution centered at the origin yields:

$$P(w_i|\alpha) = \left(\frac{1}{2}\cdot\frac{\alpha}{2}\right)\exp\left(-\frac{\alpha}{2}|w_i|\right)$$

Notice that the Laplace distribution has a sharp peak at the origin which gets higher as α is made larger.



Using the Laplace distribution as a prior for each w_i yields:

$$\Pr(w|\alpha) = \left(\frac{1}{2} \cdot \frac{\alpha}{2}\right)^d \exp\left(-\frac{\alpha}{2}\sum_i |w_i|\right)$$

The MAP estimate for w is calculated as:

$$\begin{aligned}\operatorname*{argmax}_w P(w|y) &\propto P(y|w)P(w|\alpha)\\ \ln(P(w|y)) &= \ln P(y|w) + \ln P(w|\alpha) + \text{constant}\end{aligned}$$

Which is an equivalent problem to minimizing the Lasso optimization problem.

Given X_{new} to predict Y_{new}, in the classical framework, one predicts:

$$y_{new} = w^T x_{new}$$

In the Bayesian framework, one looks at the *predictive distribution*

$$P(y_{new}|y, X, x_{new}, \sigma^2) = \int P(y_{new}|w, X, x_{new}, \sigma^2)P(w|X)\,\mathrm{d}w$$

which is Gaussian. The mean of $\Pr(y_{new}|y, X, x_{new}, \sigma^2)$ is $\tilde{w}^T x_{new}$ where $\tilde{w}$ is the posterior mean.

2.2 w and σ^2 are both given prior distributions

For the conjugate priors in this case, we use:

$$\begin{aligned}w|\sigma^2 &\sim N(\mu_0, \sigma_0^2)\\ \sigma^2 &\sim IG(\alpha, \beta)\end{aligned}$$

where IG means the inverse gamma distribution. The resulting posterior distribution is:

$$P(w, \sigma^2|y, x) \sim N(\mu_n, \sigma^2\Lambda_n^{-1})IG(\alpha_n, \beta_n)$$

where

$$\begin{aligned}\alpha_n &= \alpha + n/2\\ \beta_n &= \beta + \frac{1}{2}\left(y^T y + \mu_0^T S_0^{-1}\mu_0 - \mu_n^T \Lambda_n \mu_n\right)\\ \mu_n &= (X^T X + S_0^{-1})^{-1}(X^T y + S_0^{-1}\mu_0)\\ \Lambda_n &= X^T X + S_0^{-1}\end{aligned}$$

In practice, this model is too informative and has too many constants. People often prefer to use the *g-prior*, which is a special case of this general prior distribution. In the g-prior, the constants S_0 and μ_0 are set to:

$$\begin{aligned}S_0 &= g(X^T X)^{-1}\\ \mu_0 &= 0\end{aligned}$$

And $\alpha \to 0$, $\beta \to 0$ via limits since the inverse gamma is not defined at $\alpha = 0$, $\beta = 0$. Then, $P(\sigma^2) \propto 1/\theta^2$ is the Jeffrey's prior.

$$P(w) \sim N(0, g(X^T X)^{-1})$$

The weight random variable w has posterior mean:

$$E(w) = \frac{g}{g-1}\left(\frac{\mu_0}{g} + w_{LS}\right)$$

where w_{LS} is the least squares solution for w. Typically, setting $\mu_0 = 0$ yields:

$$\frac{g}{g+1} w_{LS}$$

Note that $X^T X$ gives some indication of which coordinates carry more relevance. This gives a very high level argument as to why having $X^T X$ might be useful within the definition of the probability distribution.

This leaves a question: How does one set the value of g? There are several strategies that can be employed:

1. Put a prior on g. This is the fully Bayesian answer. Typically, an inverse gamma distribution is used. However, at some point, one must stop going down the Bayesian rabbit hole, and one must actually assign values directly to hyper parameters.

2. *Cross-validation* can be used. In this case, a portion of the data is held out. The data which is held out is used to determine the most likely value for g.

3. *Empirical Bayes*

   $$\hat{g} = \operatorname*{argmax}_g P(y|X, g)$$

   maximizes the marginal likelihood of y after integrating out all other parameters. Typically, this is calculated numerically. While it has the disadvantage where g is choses partially based upon the input data, the Empirical Bayes remains an important and practical method.