Lec 16: Bayesian Regression

Ailin Zhang

Roadmap for Regularized Learning

- Ridge regression
- Lasso regression
- Coordinate descent
- Spline regression
- Least angle regression
- Stagewise regression / epsilon learning
- Bayesian regression
- Perceptron
- SVM
- Adaboost

Bayesian Regression

- Bayesian interpretation of regularization
  - Ridge: $\min \|\mathbf{Y} - \mathbf{X}\beta\|_{\ell_2}^2/2 + \lambda\|\beta\|_{\ell_2}^2$
  - Lasso: $\min \|\mathbf{Y} - \mathbf{X}\beta\|_{\ell_2}^2/2 + \lambda\|\beta\|_{\ell_1}$

Some Notation

- $S = \{(x_i, y_i)\}_{i=1}^{n}$ is the set of observed input/output pairs in $\mathbb{R}^d \times \mathbb{R}$ (the training set).
- X and Y denote the matrices $[x_1, \ldots, x_n]^T \in \mathbb{R}^{n \times d}$ and $[y_1, \ldots, y_n]^T \in \mathbb{R}^n$, respectively.
- β is a vector of parameters in $\mathbb{R}^p$.
- $p(Y \mid X, \beta)$ is the joint distribution over outputs Y given inputs X and the parameters.

Bayes Theorem

Theorem

$$p(\beta, \alpha) = p(\alpha \mid \beta) \cdot p(\beta)$$

- Bayesian model specifies $p(\beta, \alpha)$, usually by a measurement model, $p(\alpha \mid \beta)$ and a prior $p(\beta)$.
  - Measurement model for linear regression:

$$Y \mid X, \beta \sim \mathcal{N}\left(X\beta, \sigma_\varepsilon^2 I\right)$$

  X fixed/non-random, β is unknown.

Maximum Likelihood Estimator: ERM (Empirical risk minimization)

Measurement model:

$$Y \mid X, \beta \sim \mathcal{N}\left(X\beta, \sigma_\varepsilon^2 I\right)$$

Want to estimate β.

- Can do this without defining a prior on β.
- Maximize the likelihood, i.e. the probability of the observations.

Likelihood

- The likelihood of any fixed parameter vector β is:

$$L(\beta \mid X) = p(Y \mid X, \beta)$$

Note: we always condition on X.

ERM as a Maximum Likelihood Estimator

Measurement model:

$$Y \mid X, \beta \sim \mathcal{N}\left(X\beta, \sigma_\varepsilon^2 I\right)$$

Likelihood:

$$\begin{aligned} L(\beta \mid X) &= \mathcal{N}\left(Y; X\beta, \sigma_\varepsilon^2 I\right) \\ &\propto \exp\left(-\frac{1}{2\sigma_\varepsilon^2}\|Y - X\beta\|^2\right) \end{aligned}$$

Maximum likelihood estimator is ERM:

$$\arg\min_\beta \frac{1}{2}\|Y - X\beta\|^2$$

Bayesian Regression

Now let's consider ridge regression: Is there a probablistic model for ridge regression?

- Yes, $p(Y|X, \beta)p(\beta)$
- Measurement model:

$$Y \mid X, \beta \sim \mathcal{N}\left(X\beta, \sigma_\varepsilon^2 I\right)$$

- Add a prior

$$\beta \sim \mathcal{N}(0, I)$$

Bayesian Regression

- Take $p(Y \mid X, \beta)$ and $p(\beta)$.
- Apply Bayes' rule to get **posterior**:

$$
\begin{aligned}
p(\beta \mid X, Y) &= \frac{p(Y \mid X, \beta) \cdot p(\beta)}{p(Y \mid X)} \\
&= \frac{p(Y \mid X, \beta) \cdot p(\beta)}{\int p(Y \mid X, \beta) d\beta}
\end{aligned}
$$

- Use the posterior to estimate β.

Bayesian Estimators

- Bayes least squares estimator

  The Bayes least squares estimator for β given the observed Y is:

  $$\hat{\beta}_{BLS}(Y \mid X) = \mathbb{E}_{\beta|X,Y}[\beta]$$

  i.e. the mean of the posterior.

- Maximum a posteriori estimator

  The MAP estimator for β given the observed Y is:

  $$\hat{\beta}_{MAP}(Y \mid X) = \arg\max_{\beta} p(\beta \mid X, Y)$$

  i.e. a mode of the posterior.

Bayesian Estimators

Model:

$$Y \mid X, \theta \sim \mathcal{N}\left(X\theta, \sigma_\varepsilon^2 I\right), \quad \theta \sim \mathcal{N}(0, I)$$

Posterior:

$$\theta \mid X, Y \sim \mathcal{N}\left(\mu_{\theta|X,Y}, \Sigma_{\theta|X,Y}\right)$$

where

$$\mu_{\theta|X,Y} = X^T \left(XX^T + \sigma_\varepsilon^2 I\right)^{-1} Y$$

$$\Sigma_{\theta|X,Y} = I - X^T \left(XX^T + \sigma_\varepsilon^2 I\right)^{-1} X$$

This is Gaussian, so

$$\hat{\theta}_{MAP}(Y \mid X) = \hat{\theta}_{BLS}(Y \mid X) = X^T \left(XX^T + \sigma_\varepsilon^2 I\right)^{-1} Y$$

which corresponds to the ridge regression with $\lambda = \sigma_\epsilon^2$.

Bayesian Regression Example

Prior: $\beta \sim Laplace(\gamma)$

$p(\beta) = (\frac{\gamma}{2})^p \exp(-\gamma||\beta||_1)$

Bayesian Regression Example (Multiparameter case)

Noninformative Prior: $[\beta, \sigma^2] \sim \frac{1}{\sigma^2}$

Why Bayesian can be used for regularization?

- By placing a prior belief that the models we learn must be as simple as possible, we are able to control the complexity of the models we learn even before we learn them!
- This is exactly what we have done with our analytic derivations above: by placing a prior belief on the distribution of our model parameters (i.e. "the model parameters are normally-distributed") we are able to directly shape how complex these models are.