COMP 562 – Lecture 6

Logistic Regression -- Log-Likelihood for ± 1 Labels

Probability of a single sample is when $y \in \{-1, +1\}$:

$$p(y|\mathbf{x}, \beta_0, \beta) = \frac{1}{1 + \exp\left\{-y(\beta_0 + \mathbf{x}^T\beta)\right\}}$$

Likelihood function is:

$$\mathcal{L}(\beta_0, \beta|\mathbf{y}, \mathbf{x}) = \prod_i \frac{1}{1 + \exp\left\{-y_i(\beta_0 + \mathbf{x}_i^T\beta)\right\}}$$

Log-likelihood function is:

$$\mathcal{LL}(\beta_0, \beta|\mathbf{y}, \mathbf{x}) = -\sum_i \log\left\{1 + \exp\left\{-y_i(\beta_0 + \mathbf{x}_i^T\beta)\right\}\right\}$$

Logistic Regression -- Log-Likelihood for $0, 1$ Labels

Probability of a single sample is when $y \in \{0, 1\}$:

$$p(y|\mathbf{x}, \beta_0, \beta) = \frac{\exp\left\{y(\beta_0 + \mathbf{x}^T\beta)\right\}}{1 + \exp\left\{(\beta_0 + \mathbf{x}^T\beta)\right\}}$$

Likelihood function is:

$$\mathcal{L}(\beta_0, \beta|\mathbf{y}, \mathbf{x}) = \prod_i \frac{\exp\left\{y_i(\beta_0 + \mathbf{x}_i^T\beta)\right\}}{1 + \exp\left\{(\beta_0 + \mathbf{x}^T\beta)\right\}}$$

Log-likelihood function is:

$$\mathcal{LL}(\beta_0, \beta|\mathbf{y}, \mathbf{x}) = \sum_i y_i(\beta_0 + \mathbf{x}_i^T\beta) - \log\left\{1 + \exp\left\{(\beta_0 + \mathbf{x}_i^T\beta)\right\}\right\}$$

Ridge Penalty and Logistic Regression

Adding ridge penalty to the logistic regression achieves

1. Shrinkage of weights -- weights no longer explode in separable case
2. Even splitting between correlated weights

Ridge regularized log-likelihood for $\pm$ 1 labels:

$$\mathcal{PLL}(\beta_0, \beta|\mathbf{y}, \mathbf{x}) = -\sum_i \log\left\{1 + \exp\left\{-y_i(\beta_0 + \mathbf{x}_i^T\beta)\right\}\right\} - \frac{\lambda}{2}\|\beta\|^2$$

Ridge regularized log-likelihood for $0, 1$ labels:

$$\mathcal{PLL}(\beta_0, \beta|\mathbf{y}, \mathbf{x}) = \sum_i y_i(\beta_0 + \mathbf{x}_i^T\beta) - \log\left\{1 + \exp\left\{(\beta_0 + \mathbf{x}_i^T\beta)\right\}\right\} - \frac{\lambda}{2}\|\beta\|^2$$

Bayesian View of Penalties

We have seen two examples of supervised models

1. Linear regression, $p(y|\mathbf{x}, \beta)$ where $y \in \mathbb{R}$
2. Logistic regression, $p(y|\mathbf{x}, \beta)$ where $y \in \{-1, +1\}$

We then utilized log-likelihoods

$$\mathcal{LL}(\beta|\mathbf{y}, X) = \sum_i \log p(y_i|\mathbf{x}_i, \beta)$$

and observed that we can add penalties to log-likelihoods

$$\mathcal{LL}(\beta|\mathbf{y}, X) + \lambda f(\beta)$$

in order to deal with ill-posedness of the problems

Bayesian View of Penalties

Given a likelihood

$$p(\text{Data}|\theta)$$

Bayesian view of models treats each parameter θ as just another random variable

This random variable has a distribution called **prior** distribution

$$p(\theta)$$

Using Bayes rule we can also compute

$$\overbrace{p(\theta|\text{Data})}^{\text{posterior}} = \frac{\overbrace{p(\text{Data}|\theta)}^{\text{likelihood}}\overbrace{p(\theta)}^{\text{prior}}}{\underbrace{p(\text{Data})}_{\text{evidence}}}$$

called **posterior** distribution

Prior encodes our beliefs **before** seing the data

Posterior reflects our updated beliefs **after** seeing the data

Bayesian View of Penalties

For example we can assume a Gaussian **prior** on β_i to our linear regression model

$$\begin{aligned} \beta_i &\sim \mathcal{N}\left(0, \frac{1}{\lambda}\right), \qquad\qquad i > 0 \\ y &\sim \mathcal{N}\left(\beta_0 + \mathbf{x}^T\beta, \sigma^2\right) \end{aligned}$$

Then posterior probability of the parameter β_i:

$$p(\beta|\mathbf{y}, \mathbf{x}) = \frac{p(\mathbf{y}|\mathbf{x}, \beta)p(\beta)}{p(\mathbf{y}|\mathbf{x})}$$

Bayesian View of Penalties

We can now try to find **Maximum-A-Posteriori (MAP)** estimate of θ

$$\arg\max_{\beta} p(\beta|\mathbf{y},\mathbf{x}) = \arg\max_{\beta} \log p(\mathbf{y}|\mathbf{x},\beta) + \log p(\beta)$$

and this is equivalent to

$$\arg\max_{\beta} p(\beta|\mathbf{y},\mathbf{x}) = \arg\max_{\beta} -\sum_{i=1}^{N} \frac{1}{2\sigma^2}(y_i - \beta_0 - \mathbf{x}_i^T\beta) - \sum_{j=1}^{p} \frac{\lambda}{2}\beta_j^2 + \text{const}$$

Solving ridge regression is equivalent to finding Maximum-A-Posteriori estimate in Bayesian linear regression with Gaussian prior on weights

Softmax

Sigmoid:

$$\sigma(z) = \frac{1}{1+\exp\{-z\}} = \frac{\exp\{z\}}{1+\exp\{z\}}$$

Softmax is a generalization of sigmoid:

$$\sigma(\mathbf{z})_j = \frac{\exp\left\{z_j\right\}}{\sum_{c=1}^{C} \exp\left\{z_j\right\}}$$

For example:

$$\begin{aligned}
\sigma(\mathbf{z})_1 &= \frac{\exp\{z_1\}}{\exp\{z_1\}+\exp\{z_2\}+\exp\{z_3\}} \\
\sigma(\mathbf{z})_2 &= \frac{\exp\{z_2\}}{\exp\{z_1\}+\exp\{z_2\}+\exp\{z_3\}} \\
\sigma(\mathbf{z})_3 &= \frac{\exp\{z_3\}}{\exp\{z_1\}+\exp\{z_2\}+\exp\{z_3\}}
\end{aligned}$$

Multiclass Logistic Regression and Softmax

We can write out probability of partcular class using softmax

$$p(y = c|\mathbf{x}, \beta_0, B) = \boxed{\frac{\exp\left\{\beta_{0,c} + \mathbf{x}^T \boldsymbol{\beta}_c\right\}}{\sum_{k=1}^{C} \exp\left\{\beta_{0,k} + \mathbf{x}^T \boldsymbol{\beta}_k\right\}}}$$

where

$$B = \left[\boldsymbol{\beta}_1 \boldsymbol{\beta}_2 \ldots \boldsymbol{\beta}_C\right]$$

and each $\boldsymbol{\beta}_c$ is a vector of class specific feature weights

Note that the $p(y = c|\ \cdots)$ is a categorical distribution over C possible states, where probabilities of each state are given by softmax

Multiclass Logistic Regression -- Log-Likelihood

1. There are N samples, each in one of C classes, and p features
2. Labels are represented using one-hot vectors y_i
3. Feature matrix X contains a column of 1s -- corresponding to the bias term
4. First row of weight matrix B are bias terms
5. $\boldsymbol{\beta}_k$ is k^{th} column of matrix B

Dimensions:

- Feature matrix : $X \rightarrow N \times (p + 1)$
- Label matrix : $Y \rightarrow N \times C$
- Weight matrix : $B \rightarrow (p + 1) \times C$

Likelihood is

$$\mathcal{L}(B|Y, X) = \underbrace{\prod_{i=1}^{N}}_{\text{samples}} \underbrace{\prod_{c=1}^{C}}_{\text{classes}} \left[\frac{\exp\left\{\mathbf{x}_i^T \boldsymbol{\beta}_c\right\}}{\sum_{k=1}^{C} \exp\left\{\mathbf{x}_i^T \boldsymbol{\beta}_k\right\}}\right]^{y_{i,c}}$$

Log-likelihood is

$$\mathcal{LL}(\beta_0, B|Y, X) = \sum_{i=1}^{N} \sum_{c=1}^{C} y_{i,c} \left(\mathbf{x}_i^T \boldsymbol{\beta}_c - \log\left\{\sum_{k=1}^{C} \exp\left\{\mathbf{x}_i^T \boldsymbol{\beta}_k\right\}\right\}\right)$$

Multiclass Logistic Regression -- Regularized Log-Likelihood

Ridge regularized log-likelihood

$$\mathcal{PLL}(B|Y,X) = \sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\left(\mathbf{x}_i^T\boldsymbol{\beta}_c - \log\left\{\sum_{k=1}^{C}\exp\left\{\mathbf{x}_i^T\boldsymbol{\beta}_k\right\}\right\}\right) - \frac{\lambda}{2}\sum_{k=1}^{C}\sum_{j=1}^{p}\beta_{j,k}^2$$

Note that we keep the last column of B fixed at 0 to get rid of excess parameters

These parameters will not contribute to the regularization -- sum of their squares is 0

Cross-Entropy

Frequently you will encounter mentions of cross-entropy. It is negative log likelihood of multiclass logistic

$$\begin{aligned}\text{crossentropy}(B) &= -\mathcal{LL}(B|Y,X)\\ &= -\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\left(\mathbf{x}_i^T\boldsymbol{\beta}_c - \log\left\{\sum_{k=1}^{C}\exp\left\{\mathbf{x}_i^T\boldsymbol{\beta}_k\right\}\right\}\right)\end{aligned}$$

Ridge regularized cross-entropy

$$\text{crossentropy}(B) = -\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\left(\mathbf{x}_i^T\boldsymbol{\beta}_c - \log\left\{\sum_{k=1}^{C}\exp\left\{\mathbf{x}_i^T\boldsymbol{\beta}_k\right\}\right\}\right) + \frac{\lambda}{2}\sum_{k=1}^{C}\sum_{j=1}^{p}\beta_{j,k}^2$$

Note the sign flip in the regularization