National Technical University of Ukraine
“Igor Sikorsky Kyiv Polytechnic Institute”
Institute of Physics and Technology

Lecture 14

Objects and Systems Identification Methods. Logistic Regression

Dmytro Progonov,

PhD, Associate Professor,

Department Of Physics and Information Security Systems

Content

- Model specification;
- Advantages of logistic regression;
- Model fitting:
  - Maximum Likelihood Estimation;
  - Steepest descent;
  - Newton’s method;
  - Iteratively reweighted least squares;
  - Quasi-Newton (variable metric) methods;
  - ℓ_2 regularization.

Model specification

We can generalized linear regression to the (binary) classification setting by making *two changes*. *First* we replace the Gaussian distribution for y with a Bernoulli distribution:

$$\boxed{p(y|\mathbf{x},\boldsymbol{\theta}) = Ber\big(y|\mu(\mathbf{x})\big)}$$

where $\mu(\mathbf{x}) = \mathbb{E}[y|\mathbf{x}] = p(y = 1|\mathbf{x})$.

Second we compute a linear combination of the inputs, as before, but then we pass this through a function that ensure $0 \leq \mu(\mathbf{x}) \leq 1$:

$$\boxed{\mu(\mathbf{x}) = sigm(\mathbf{w}^T\mathbf{x})}$$

where $sigm(\cdot)$ – sigmoid (logistic or logit) function.

$$\boxed{sigm(\eta) \triangleq \frac{1}{1 + \exp[-\eta]} = \frac{e^{\eta}}{e^{\eta} + 1}}$$

Then final form of logistic regression model is:

$$\boxed{p(y|\mathbf{x},\mathbf{w}) = Ber(y|sigm(\mathbf{w}^T\mathbf{x})\ )}$$

Advantages of logistic regression

- *Logistic regression (LR) are easy to fit* – we mean that the algorithms are simple to implement, and are very fast (even linear time in the number of non-zeros in the dataset);
- *LR models are easy to interpret*;
- *LR model are easy to extend to multi-class classification*;
- *LR model can be easily be extended to handle non-linear decision boundaries* by using kernels or by learning features from data.

Model fitting. Maximum Likelihood Estimation

The ***negative log-likelihood*** for logistic regression is given by:

$$\boxed{NLL(\mathbf{w}) = -\sum_{i=1}^{N} \log\left[\mu_i{}^{\mathbb{I}(y_i=1)} \times (1-\mu_i)^{\mathbb{I}(y_i=0)}\right] = -\sum_{i=1}^{N} [y_i \log[\mu_i] + (1-y_i)\log[1-\mu_i]]}$$

where $\mu_i = sigm(\mathbf{w}^T\mathbf{x}_i)$. This is also called ***cross-entropy function***.

Unlike linear regression, we can no longer write down the MLE in closed form. Instead, we need to use an optimization algorithm to compute it. For this we need to derive the gradient and Hessian.

$$\boxed{\mathbf{g} = \frac{d}{d\mathbf{w}} NLL(\mathbf{w}) = \sum_i (\mu_i - y_i)\mathbf{x}_i = \mathbf{X}^T(\boldsymbol{\mu} - \mathbf{y})}$$

$$\boxed{\mathbf{H} = \frac{d}{d\mathbf{w}} \mathbf{g}(\mathbf{w})^T = \sum_i (\nabla_{\mathbf{w}}\mu_i)\mathbf{x}_i{}^T = \sum_i \mu_i(-\mu_i)\,\mathbf{x}_i\mathbf{x}_i{}^T}$$

Model fitting. Steepest descent

Perhaps the simplest algorithm for unconstrained optimization is ***gradient descent***, also known as ***steepest descent***:

$$\boxed{\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta_k \mathbf{g}_k}$$

where η_k is the ***step size*** or ***learning rate***.

The common approach to determine the step size is usage of ***line minimization*** or ***line search algorithm***:

$$\boxed{\eta_{optim} = \operatorname*{argmin}_{\eta}(\boldsymbol{\theta}_k + \eta \mathbf{g}_k)}$$

To suppress the *zig-zag behavior*, we use a simple heuristic – add a momentum term:

$$\boxed{\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \eta_k \mathbf{g}_k + \mu_k(\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k-1})}$$

where μ_k ($\mu_k \in [0; 1]$) control the importance of momentum term.

Model fitting. Newton’s method

1. Initialize $\boldsymbol{\theta}_0$;
2. For $k = 1,2,\cdots$ until convergence do:
    1. Evaluate $\mathbf{g}_k = \nabla f(\boldsymbol{\theta}_k)$;
    2. Evaluate $\mathbf{H}_k = \nabla^2 f(\boldsymbol{\theta}_k)$;
    3. Solve $\mathbf{H}_k \mathbf{d}_k = (-\mathbf{g}_k)$ for $\mathbf{d}_k$;
    4. Use line search to find step size η_k along $\mathbf{d}_k$;
    5. Update $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \eta_k \mathbf{d}_k$.

Newton’s method called ***second order*** optimization method.

Model fitting.
Iteratively reweighted least squares

```
1. Initialize $\mathbf{w} = \mathbf{0}_D$;
2. Evaluate $w_0 = \log[\bar{y}/(1 - \bar{y})]$;
3. Until converged do:
    1. $\eta_i = w_0 + \mathbf{w}^T \mathbf{x}_i$;
    2. $\mu_i = sigm(\eta_i)$;
    3. $s_i = \mu_i(1 - \mu_i)$;
    4. $z_i = \eta_i + \frac{y_i - \mu_i}{s_i}$;
    5. $\mathbf{S} = diag(s_{1:N})$;
    6. $\mathbf{w} = (\mathbf{X}^T \mathbf{S} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{S} \mathbf{z}$.
```

Model fitting.
Quasi-Newton (variable metric) methods

Limitation of practical usage the Newton’s method is ***necessary to compute the Hessian*** that can be expensive for some cases.

Quasi-Newton methods iteratively ***build up an approximation to the Hessian*** using information gleaned from the gradient vector at each step (Broyden, Fletcher, Goldfarb and Shanno):

$$\boxed{\mathbf{H}_{k+1} \approx \mathbf{B}_{k+1} = \mathbf{B}_k + \frac{\mathbf{y}_k {\mathbf{y}_k}^T}{{\mathbf{y}_k}^T \mathbf{s}_k} - \frac{(\mathbf{B}_k \mathbf{s}_k)(\mathbf{B}_k \mathbf{s}_k)^T}{{\mathbf{s}_k}^T \mathbf{B}_k \mathbf{s}_k}}$$

$$\boxed{\mathbf{s}_k = \boldsymbol{\theta}_k - \boldsymbol{\theta}_{k-1}}$$

$$\boxed{\mathbf{y}_k = \mathbf{g}_k - \mathbf{g}_{k-1}}$$

Model fitting. ℓ_2 regularization

Regularization is important in the classification setting even if we have lots of data. To see why, suppose the data is linearly separable. On this case, the MLE is obtained when $||\mathbf{w}|| \rightarrow \infty$, corresponding to an infinitely steep sigmoid function, also known as linear ***threshold unit***.

To prevent this we can use ℓ_2 regularization, just we did with ridge regression:

$$\boxed{\acute{f}(\mathbf{w}) = NLL(\mathbf{w}) + \lambda \mathbf{w}^T \mathbf{w}}$$

$$\boxed{\acute{\mathbf{g}}(\mathbf{w}) = \mathbf{g}(\mathbf{w}) + 2\lambda \mathbf{w}}$$

$$\boxed{\acute{\mathbf{H}}(\mathbf{w}) = \mathbf{H}(\mathbf{w}) + 2\lambda \mathbf{I}}$$

Conclusion

- Model specification for Logistic regression was shown;
- Common and special methods for fitting the Logistic regression were considered.