COMPSCI 689
Lecture 5: MAP Estimation and Logistic Regression

Benjamin M. Marlin

College of Information and Computer Sciences
University of Massachusetts Amherst

Slides by Benjamin M. Marlin (marlin@cs.umass.edu).

Logistic Regression

- Suppose $y \in \{-1, 1\}$ and $\mathbf{x} \in \mathbb{R}^D$.
- Let $\theta = [\mathbf{w}, b]$ where $\mathbf{w} \in \mathbb{R}^D$, $\mathbf{b} \in \mathbb{R}$.
- $\mathbf{w}$ are referred to as the weights, b is the bias.
- Linear logistic regression has the form:

$$P(\mathbf{Y} = y|\mathbf{X} = x, \theta) = \frac{1}{1 + \exp(-y(\mathbf{w}\mathbf{x}^T + b))}$$

$$= \left(\frac{1}{1 + \exp(-(\mathbf{w}\mathbf{x}^T + b))}\right)^{[y=1]} \left(\frac{\exp(-(\mathbf{w}\mathbf{x}^T + b))}{1 + \exp(-(\mathbf{w}\mathbf{x}^T + b))}\right)^{[y=-1]}$$

Logistic Regression with Bias Absorption

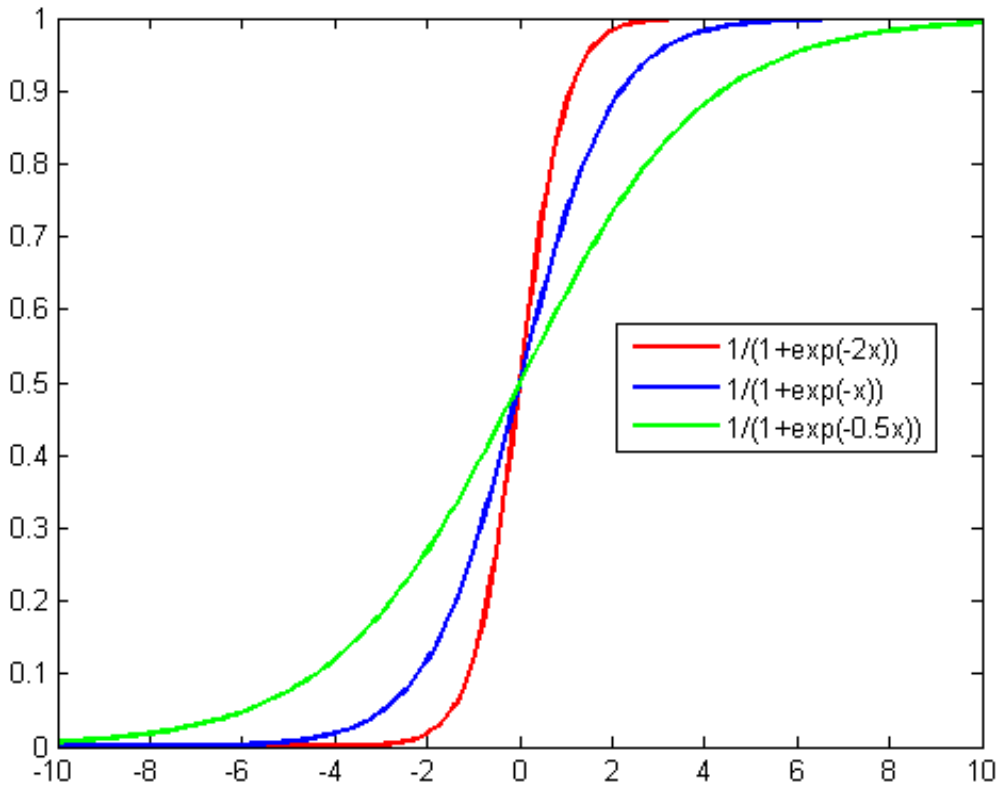
- We can again use bias absorption to incorporate b into $\mathbf{w}$:

$$P(\mathbf{Y} = y|\mathbf{X} = x, \theta) = \frac{1}{1 + \exp(-y\mathbf{w}\mathbf{x}^T)}$$

$$= \left(\frac{1}{1 + \exp(-\mathbf{w}\mathbf{x}^T)}\right)^{[y=1]} \left(\frac{\exp(-\mathbf{w}\mathbf{x}^T)}{1 + \exp(-\mathbf{w}\mathbf{x}^T)}\right)^{[y=-1]}$$

Logistic Function

$$f(x) = \frac{1}{1 + \exp(-ax)}$$



Multiclass Logistic Regression

- Linear logistic regression can also be extended to the multiclass case where $y \in \{1, ..., C\}$:

$$P(Y = y|\mathbf{x}, \theta) = \prod_{c=1}^{C} \left(\frac{\exp(\mathbf{w}_c \mathbf{x}^T)}{\sum_{c' \in \mathcal{Y}} \exp(\mathbf{w}_{c'} \mathbf{x}^T)} \right)^{[y=c]}$$

- There is one weight vector $\mathbf{w}_c$ per class (again assuming bias absorption).
- Note that this parameterization is actually redundant due to the normalization constraint. This redundancy can be removed by asserting that $\mathbf{w}_c = 0$ for one of the C classes.

MLE for Logistic Regression

- Under the MLE framework, the logistic regression model parameters $\theta = \{(\mathbf{w}_c), c \in \mathcal{Y}\}$ are selected to optimize the conditional log likelihood given a data set $\mathcal{D} = \{(y_n, \mathbf{x}_n), n = 1 : N\}$:

$$\theta_* = \arg\max_\theta l(\mathcal{D}, \theta) = \arg\max_\theta \sum_{n=1}^{N} \log P(Y = y_n | \mathbf{X} = \mathbf{x}_n, \theta)$$

$$= \arg\max_\theta \sum_{n=1}^{N} \left[\left(\sum_{c=1}^{C} [y_n = c] \mathbf{w}_c \mathbf{x}_n^T \right) - \log \left(\sum_{c' \in \mathcal{Y}} \exp(\mathbf{w}_{c'} \mathbf{x}_n^T) \right) \right]$$

Gradient for Logistic Regression

- The gradient of the log likelihood function is given below:

$$\nabla_{\mathbf{w}_i} l(\mathcal{D}, \theta) = \sum_{n=1}^{N} \Big([y_n = i] - P(Y = i | \mathbf{X} = \mathbf{x}_n, \theta) \Big) \mathbf{x}_n$$

- It turns out that the conditional log likelihood function for this model is convex, but the gradient equation $\nabla l(\mathcal{D}, \theta) = 0$ has no analytic solution. We learn the model using numerical optimization.

Prediction

- What value should we predict given the learned parameters θ_*?
- In classification, the standard loss function is the classification error or zero-one loss $L_{0/1}[y, y'] = [y \neq y']$.
- Under the expected loss minimization framework we obtain the result:

$$f_*(x) = \arg\max_{y \in \mathcal{Y}} P_*(Y = y | X = x)$$

- Using $P(Y|X, \theta_*)$ to approximate $P_*(Y|X)$, we have:

$$f_*(x) \approx \arg\max_{y \in \mathcal{Y}} P(Y = y | X = x, \theta_*)$$

- In other words, we predict the most probable value of Y.

What to Do with Small N?

- If we have an idea of what the optimal parameter values should be, we can use this information to help deal with the problem of having low sample size.
- Formally, assume we have a parametric probability distribution over the parameters $P(\theta|\theta_0)$ that we formed before we had access to data. This is called the prior distribution.
- Given a sample of data $\mathcal{D}$ and the likelihood function $P(\mathcal{D}|\theta)$, we can compute an updated distribution over the parameters θ using Bayes Rule. This is called the posterior distribution of θ:

$$P(\theta|\mathcal{D},\theta_0) = \frac{P(\mathcal{D}|\theta)P(\theta|\theta_0)}{\int_\Theta P(\mathcal{D}|\theta)P(\theta|\theta_0)d\theta}$$

MAP Estimation

- An alternative to picking θ_* by maximizing the likelihood is to pick θ_* to maximize the posterior distribution of θ. This is called *maximum a posteriori estimation* or MAP estimation:

$$\begin{aligned}\theta_* &= \arg\max_{\theta \in \Theta} P(\theta|\mathcal{D}, \theta_0) \\ &= \arg\max_{\theta \in \Theta} \sum_{n=1}^{N} \log P(Y = y_n|\mathbf{x}_n, \theta) + \log P(\theta|\theta_0)\end{aligned}$$

MAP Estimation: Properties

- Note that the solution of this optimization problem converges to the MLE as N goes to infinity since the data will overwhelm any non-degenerate prior in the limit.
- On the other hand, when N is small, the influence of the prior can help to produce much better parameter estimates when the prior $P(\theta|\theta_0)$ is reasonable.

A Prior for Weight Vectors

- To apply MAP estimation to a linear model, we need a prior distribution on $\mathbf{w}$.
- A common prior distribution for a real-valued parameter vector is a spherical Gaussian distribution centered at 0: $P(\mathbf{w}|\tau) = \mathcal{N}(\mathbf{w}; 0, \tau^2 I)$.
- The parameter τ^2 is the marginal variance of the prior distribution.
- Note that exactly the same construction works for both linear and logistic regression.

MAP for Linear Regression

The derivation is similar to finding the MLE. The final result is:

$$\mathbf{w}_* = (\mathbf{X}^T\mathbf{X} + \frac{\sigma^2}{\tau^2}I)^{-1}\mathbf{X}^T\mathbf{y}$$

This model is also known as ridge regression. It is often better to use some $\tau > 0$ to limit the complexity of w. This also makes the estimates for $\mathbf{w}_*$ numerically much more stable.

It's common to parameterize this model using $\lambda = \frac{\sigma^2}{\tau^2}$ to again avoid the need to estimate σ.

MAP for Logistic Regression

- If we assume a zero-mean spherical Gaussian prior on $\mathbf{w}_c$ for each c, we arrive at the following optimization problem whose solution is the MAP estimate of the parameters:

$$\theta_* = \arg\max_{\theta} \sum_{n=1}^{N} \log P(Y = y_n | \mathbf{X} = \mathbf{x}_n, \theta) - \lambda \sum_{c \in \mathcal{Y}} \|\mathbf{w}_c\|_2^2$$

- The gradient of this objective is nearly identical to the case with no prior, but It still can't be solved analytically...