COMPSCI 689
Lecture 7: Generalizing Linear Models

Benjamin M. Marlin

College of Information and Computer Sciences
University of Massachusetts Amherst

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

GLMs

- We have seen three examples of models based on linear functions: linear Gaussian models and binary and multi-class logistic regression.
- All three models can be can be thought of as mapping linear functions of the data into the parameter space of a simple, unconditional probability model on Y.
- These are all special cases of a very flexible modeling framework known as generalized linear models (GLMs).

GLMs

- A conditional probability model is a GLM if $P(Y|\mathbf{X} = \mathbf{x}, \mathbf{w})$ has the property:

$$\mathbb{E}[Y|\mathbf{x}] = \mu(\mathbf{x}) = g^{-1}(\mathbf{w}\mathbf{x}^T)$$

- The function g is called the link function, and g^{-1} is the inverse function of g (i.e.: $g^{-1}(g(z)) = z$), also called the mean function.

GLMs

- The linear Gaussian regression model is a GLM constructed from the normal distribution.
- Binary logistic regression is a GLM constructed from the Bernoulli distribution.
- Multi-class logistic regression is a GLM constructed from the Multinoulli (general categorical) distribution.
- It's easy to construct other GLMs to model different output spaces $\mathcal{Y}$ using other basic probability distributions that are parameterized in terms of their means.

Example: Exponential GLM

- Suppose we have a process that produces data such that $y \in \mathbb{R}^{\geq 0}$ and $\mathbf{x} \in \mathbb{R}^D$.
- One distribution that matches the support of y is the exponential distribution:

$$P(Y = y|\beta) = \frac{1}{\beta} \exp(-\frac{y}{\beta})$$

- The model satisfies $\mathbb{E}[Y] = \beta$, and it is required that $\beta > 0$.
- Suppose we want to model β using a GLM where $\beta = g^{-1}(\mathbf{w}\mathbf{x}^T)$.
- What function should we choose for g^{-1}?

Example: Exponential GLM

- Since $\beta > 0$, we need that $g^{-1} : \mathbb{R} \to \mathbb{R}^{>0}$.
- One function that satisfies this property is $g^{-1} = \exp$.
- This gives us the model:

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

- How can we learn the parameters $\mathbf{w}$?

Example: Exponential GLM MLE

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

$$\begin{aligned}
\mathbf{w}_* &= \arg\max_{\mathbf{w}} l(\mathcal{D}, \mathbf{w}) \\
&= \arg\max_{\mathbf{w}} \sum_{n=1}^{N} \log P(Y = y_n | \mathbf{X} = \mathbf{x}_n, \mathbf{w}) \\
&= \arg\max_{\mathbf{w}} \sum_{n=1}^{N} \left(-\mathbf{w}\mathbf{x}_n^T - y_n \exp(-\mathbf{w}\mathbf{x}_n^T) \right)
\end{aligned}$$

Example: Exponential GLM Gradient

The gradient of the log likelihood is given by:

$$\nabla l(\mathcal{D}, \mathbf{w}) = \sum_{n=1}^{N} \left(-1 + y_n \exp(-\mathbf{w}\mathbf{x}_n^T)\right) \mathbf{x}_n^T$$

The gradient system $\nabla l(\mathcal{D}, \mathbf{w}) = 0$ can not be solved analytically for this model, so numerical optimization methods must be used.

Example: Poisson GLM

- Suppose we have a process that produces data such that $y \in \mathbb{Z}^{\geq 0}$ and $\mathbf{x} \in \mathbb{R}^D$.
- One distribution that matches the support of y is the poisson distribution:

$$P(Y = y|\lambda) = \frac{\lambda^y \exp(-\lambda)}{y!}$$

- The model satisfies $\mathbb{E}[Y] = \lambda$, and it is required that $\lambda > 0$.
- We can again use $\lambda = \exp(\mathbf{w}\mathbf{x}^T)$.

Example: Poisson GLM

- This gives us the model:

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

- How can we learn the parameters $\mathbf{w}$?

$$\begin{aligned}
\mathbf{w}_* &= \arg\max_{\mathbf{w}} l(\mathcal{D}, \mathbf{w}) \\
&= \arg\max_{\mathbf{w}} \sum_{n=1}^{N} \log P(Y = y_n|\mathbf{X} = \mathbf{x}_n, \mathbf{w}) \\
&= \arg\max_{\mathbf{w}} \sum_{n=1}^{N} \left(y_n \mathbf{w}\mathbf{x}_n^T - \exp(\mathbf{w}\mathbf{x}_n^T) - \log(y!)\right)
\end{aligned}$$

Example: Poisson GLM Gradient

The gradient of the log likelihood is given by:

$$\nabla l(\mathcal{D}, \mathbf{w}) = \sum_{n=1}^{N} \left(y_n - \exp(\mathbf{w}\mathbf{x}_n^T)\right) \mathbf{x}_n^T$$

The gradient system $\nabla l(\mathcal{D}, \mathbf{w}) = 0$ can not be solved analytically for this model, so numerical optimization methods must be used.

The Exponential Family

- It is sometimes useful to build GLMs using the exponential family form of a probability distribution $P(Y = y)$ when the distribution admits such a form:

$$P(Y = y|\theta) = h(y) \exp\left(\theta\phi(y)^T - A(\theta)\right)$$

- $\phi(y) \in \mathbb{R}^D$ is called the vector of sufficient statistics.
- θ are referred to as the natural parameters of the distribution.
- $A(\theta)$ is referred to as the log partition function or the cumulant function. It has the useful property $\frac{\partial A(\theta)}{\partial \theta} = \mathbb{E}[\phi(y)] = \mu$.
- $h(y)$ is a scaling constant, often equal to 1.

The Exponential Family

- Single-parameter exponential family distributions admit an invertible mapping between the natural parameters θ and the mean parameters $\mathbb{E}[\phi(y)] = \mu$: $\psi(\mu) = \theta$.
- This mapping is uniquely determined by the distribution.
- Most basic distributions you are familiar with are members of the exponential family and their exponential family forms can be inferred from their more common textbook definitions.
- Since all Exponential Family distributions can be written in the same form, we can obtain generalized results for procedures like computing the MLE of θ.

Exponential Family and GLMs

- To form a GLM from the exponential family form of a distribution, we need to again specify a mean function of the form $\mu(\mathbf{x}) = \mathbb{E}[y|\mathbf{x}] = g^{-1}(\mathbf{w}\mathbf{x}^T)$.
- The so called *canonical link function* is given by $g = \psi$, where ψ maps from the mean μ parameters to the natural parameters θ.
- We thus have: $\mathbb{E}[y|\mathbf{x}] = \mu(\mathbf{x}) = \psi^{-1}(\mathbf{w}\mathbf{x}^T)$.
- However, an exponential family distribution is defined in terms of θ, not μ, but we can map between them using $\theta = \psi(\mu)$.
- The canonical link function thus yields:

$$\theta(\mathbf{x}) = \psi(\mu(\mathbf{x})) = \psi(\psi^{-1}(\mathbf{w}\mathbf{x}^T)) = \mathbf{w}\mathbf{x}^T$$

Exponential Family and GLMs

- Assuming the canonical link yields the GLM model:

$$P(Y = y|\mathbf{X} = \mathbf{x}, \mathbf{w}) = h(y) \exp\left(\mathbf{w}\mathbf{x}^T \phi(y) - A(\mathbf{w}\mathbf{x}^T)\right)$$

- Any GLM of this form has a gradient of the log likelihood function equal to:

$$\nabla l(\mathcal{D}, \mathbf{w}) = \sum_{n=1}^{N} (\phi(y_n) - \mu(\mathbf{x}_n))\mathbf{x}_n = \sum_{n=1}^{N} (\phi(y_n) - \psi^{-1}(\mathbf{w}\mathbf{x}_n^T))\mathbf{x}_n$$

- Note that the canonical link function doesn't have to be used, and that in some cases it can result in parameter values that aren't actually valid (for example, if the parameters have constraints).

The Problem with Linear Models

- The problem with generalized linear models is that the relationship between the expectation of Y and $\mathbf{x}$ is quite constrained.
- Since g must be invertible, $g^{-1}(\mathbf{w}\mathbf{x}^T)$ will be a monotonic function of $\mathbf{w}\mathbf{x}^T$.
- This makes the regression functions learned by the linear Gaussian model and the class boundaries learned by logistic regression linear in $\mathbf{x}$.
- We can make GLMs more flexible using basis functions expansions.

Basis Function Expansion

- A simple solution to the linearity problem is to apply a set of functions ϕ_1,...,ϕ_K to the raw feature vector $\mathbf{x}$ to map it in to a new feature space:

$$\phi(\mathbf{x}) = [\phi_1(\mathbf{x}), ..., \phi_K(\mathbf{x})]$$

- This is called a *basis function expansion* since $K > D$ in general. This requires that we know the functions ϕ_1,...,ϕ_K that we want to apply in advance.

- We then define a generalized linear model in this new feature space using:

$$\mathbb{E}[Y] = g^{-1}(\mathbf{w}\phi(\mathbf{x})^T)$$

Basis Function Expansion Examples

- **Univariate Functions:** We can set $\phi_k(\mathbf{x})$ to any univariate function of a single x_d to obtain mappings like $\phi(\mathbf{x}) = [x_1, x_2, \sin(x_1), \exp(x_2)]$, etc.
- **Degree 2 Polynomial Basis:** We include all single features x_d, their squares x_d^2, and all products of two distinct features $x_d x_{d'}$.
- **Degree B Polynomial Basis:** We include all single features x_d, and all unique products of between 2 and B features.

Properties of Basis Function Expansions

- A basis expansion can be used to break the monotonic relationship between $\mathbb{E}[Y|\mathbf{x}]$ imposed by a GLM.
- If we have a GLM with a convex objective function, the objective function remains convex under a basis expansion.
- If we have an algorithm implementing maximum likelihood learning for a GLM, we can use it to learn the parameters of the GLM under an arbitrary basis expansion. We just need to modify the data we feed to the learning algorithm.
- However, basis expansions have two significant drawbacks: you need to know the ϕ_k functions in advance, and the computational complexity of learning depends on the dimensionality K of the basis expansion, not the original data dimension D.

Summary

- We now have an extremely flexible framework for estimating conditional probability models of the form $P(Y|\mathbf{X} = \mathbf{x}, \theta)$.
- We can build models for any output space $\mathcal{Y}$ by selecting an unconditional probability distribution with the corresponding support and building a GLM from it.
- To deal with more complex relationships between $\mathbf{x}$ and $E[Y|\mathbf{x}]$, we can introduce the use of basis expansions.
- We can learn models using maximum likelihood, resorting to numerical optimization when needed.
- To control the complexity of the models in low-data settings, we can apply MAP estimation or (equivalently) parameter regularization.