6

Conditional Densities

A number of machine learning algorithms can be derived by using conditional exponential families of distribution (Section 2.3). Assume that the training set $\{(x_1, y_1), \ldots, (x_m, y_m)\}$ was drawn iid from some underlying distribution. Using Bayes rule (1.15) one can write the likelihood

$$p(\theta|X,Y) \propto p(\theta)p(Y|X,\theta) = p(\theta)\prod_{i=1}^{m} p(y_i|x_i,\theta), \tag{6.1}$$

and hence the negative log-likelihood

$$-\log p(\theta|X,Y) = -\sum_{i=1}^{m} \log p(y_i|x_i,\theta) - \log p(\theta) + \text{const.} \tag{6.2}$$

Because we do not have any prior knowledge about the data, we choose a zero mean unit variance isotropic normal distribution for $p(\theta)$. This yields

$$-\log p(\theta|X,Y) = \frac{1}{2}\|\theta\|^2 - \sum_{i=1}^{m} \log p(y_i|x_i,\theta) + \text{const.} \tag{6.3}$$

Finally, if we assume a conditional exponential family model for $p(y|x,\theta)$, that is,

$$p(y|x,\theta) = \exp\left(\langle\phi(x,y),\theta\rangle - g(\theta|x)\right), \tag{6.4}$$

then

$$-\log p(\theta|X,Y) = \frac{1}{2}\|\theta\|^2 + \sum_{i=1}^{m} g(\theta|x_i) - \langle\phi(x_i,y_i),\theta\rangle + \text{const.} \tag{6.5}$$

where

$$g(\theta|x) = \log\sum_{y\in\mathcal{Y}} \exp\left(\langle\phi(x,y),\theta\rangle\right), \tag{6.6}$$

is the log-partition function. Clearly, (6.5) is a smooth convex objective function, and algorithms for unconstrained minimization from Chapter 5 can be used to obtain the maximum aposteriori (MAP) estimate for θ. Given the optimal θ, the class label at any given x can be predicted using

$$y^* = \operatorname*{argmax}_y p(y|x, \theta). \tag{6.7}$$

In this chapter we will discuss a number of these algorithms that can be derived by specializing the above setup. Our discussion unifies seemingly disparate algorithms, which are often discussed separately in literature.

6.1 Logistic Regression

We begin with the simplest case namely binary classification[1]. The key observation here is that the labels $y \in \{\pm 1\}$ and hence

$$g(\theta|x) = \log\left(\exp\left(\langle \phi(x, +1), \theta\rangle\right) + \exp\left(\langle \phi(x, -1), \theta\rangle\right)\right). \tag{6.8}$$

Define $\hat{\phi}(x) := \phi(x, +1) - \phi(x, -1)$. Plugging (6.8) into (6.4), using the definition of $\hat{\phi}$ and rearranging

$$p(y = +1|x, \theta) = \frac{1}{1 + \exp\left(\left\langle -\hat{\phi}(x), \theta\right\rangle\right)} \text{ and}$$
$$p(y = -1|x, \theta) = \frac{1}{1 + \exp\left(\left\langle \hat{\phi}(x), \theta\right\rangle\right)},$$

or more compactly

$$p(y|x, \theta) = \frac{1}{1 + \exp\left(\left\langle -y\hat{\phi}(x), \theta\right\rangle\right)}.$$

Since $p(y|x, \theta)$ is a logistic function, hence the name logistic regression. The classification rule (6.7) in this case specializes as follows: predict $+1$ whenever $p(y = +1|x, \theta) \geq p(y = -1|x, \theta)$ otherwise predict -1. However

$$\log \frac{p(y = +1|x, \theta)}{p(y = -1|x, \theta)} = \left\langle \hat{\phi}(x), \theta\right\rangle,$$

therefore one can equivalently use $\operatorname{sign}\left(\left\langle \hat{\phi}(x), \theta\right\rangle\right)$ as our prediction function. Next we turn our attention to deriving the log-likelihood. After some simple algebraic manipulation one can write

$$g(\theta|x) - \langle \phi(x, +1), \theta\rangle = \log\left(1 + \exp\left(\left\langle \hat{\phi}(x), \theta\right\rangle\right)\right) - \left\langle \hat{\phi}(x), \theta\right\rangle \text{ and}$$
$$g(\theta|x) - \langle \phi(x, -1), \theta\rangle = \log\left(1 + \exp\left(\left\langle -\hat{\phi}(x), \theta\right\rangle\right)\right) + \left\langle \hat{\phi}(x), \theta\right\rangle.$$

[1] The name logistic *regression* is a misnomer!

The log-likelihood (6.5) can now be written compactly by combining the above two equations as

$$\frac{1}{2}\|\theta\|^2 + \sum_{i=1}^{m} \log\left(1 + \exp\left(\left\langle y_i\hat{\phi}(x_i), \theta\right\rangle\right)\right) - y_i\left\langle \hat{\phi}(x_i), \theta\right\rangle + \text{const.}$$

To minimize the above objective function we first compute the gradient.

$$\begin{aligned}\nabla J(\theta) &= \theta + \sum_{i=1}^{m} \frac{\exp\left(\left\langle y_i\hat{\phi}(x_i), \theta\right\rangle\right)}{1 + \exp\left(\left\langle y_i\hat{\phi}(x_i), \theta\right\rangle\right)} y_i\hat{\phi}(x_i) - y_i\hat{\phi}(x_i) \\ &= \theta + \sum_{i=1}^{m} (p(y_i|x_i, \theta) - 1) y_i\hat{\phi}(x_i).\end{aligned}$$

Notice that the second term of the gradient vanishes whenever $p(y_i|x_i, \theta) = 1$. Therefore, one way to interpret logistic regression is to view it as a method to maximize $p(y_i|x_i, \theta)$ for each point (x_i, y_i) in the training set. Since the objective function of logistic regression is twice differentiable one can also compute its Hessian

$$\nabla^2 J(\theta) = -\sum_{i=1}^{m} p(y_i|x_i, \theta)(1 - p(y_i|x_i, \theta))\hat{\phi}(x_i)\hat{\phi}(x_i)^\top,$$

where we used $y_i^2 = 1$. The Hessian can be used in the Newton method (Section 5.2.6) to obtain the optimal parameter θ.

6.2 Regression

6.2.1 Conditionally Normal Models

fixed variance

6.2.2 Posterior Distribution

integrating out vs. Laplace approximation, efficient estimation (sparse greedy)

6.2.3 Heteroscedastic Estimation

explain that we have two parameters. not too many details (do that as an assignment).