CS142: Machine Learning Spring 2017

Lecture 4

Instructor: Pedro Felzenszwalb *Scribes: Dan Xiang, Tyler Dae Devlin*

Validation

Given a set of data $A = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, we can partition the data into a training set T and a validation set V. We build the model using only the data in T and test this classifier on V. This process is called *validation* and is a way to guard against overfitting.

Least squares regression and maximum likelihood estimation

When we set up the regression problem last lecture, our goal was to minimize the sum-of-squares error

$$E(w) = \frac{1}{2} \sum_{i=1}^{n} (f_w(x_i) - y_i)^2.$$

In this lecture, we will justify this particular choice of an error function.

Recall that $f_w(x) = w \cdot \phi(x)$, where each of w, x, and $\phi(x)$ is a vector. Assume that the data $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ are generated by the true target function $f_{\hat{w}}(x)$ where $\hat{w} \in \mathbb{R}^m$ is unknown. Furthermore, assume that there is some random Gaussian noise associated with each y_i. More precisely, $y_i = f_{\hat{w}}(x_i) + \epsilon_i$ where $\{\epsilon_i\}_{i=1}^n$ are mutually independent with

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2),$$

where $\sigma \in (0, \infty)$ is unknown and is the same for all $i = 1, \ldots, n$. Then we have

$$y_i \sim \mathcal{N}(f_{\hat{w}}(x_i), \sigma^2).$$

We now introduce the notion of the likelihood function $L : \mathbb{R}^m \to \mathbb{R}$. L assigns to each $w \in \mathbb{R}^m$ a real number representing the likelihood of the given w. The likelihood function is defined by

$$L(w) \doteq P(y_1, \ldots, y_n | x_1, \ldots, x_n, w).$$

The maximum likelihood estimate of w, denoted w^*, is

$$w^* = \underset{w \in \mathbb{R}^m}{\operatorname{argmax}} L(w).$$

Assuming conditional independence of the output values $(y_1, \ldots, y_n)$ given the features $(x_1, \ldots, x_n)$, the above is equal to

$$\begin{aligned} \underset{w \in \mathbb{R}^m}{\operatorname{argmax}} L(w) &= \underset{w \in \mathbb{R}^m}{\operatorname{argmax}} \prod_{i=1}^{n} P(y_i | x_i, w) \\ &= \underset{w \in \mathbb{R}^m}{\operatorname{argmax}} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(y_i - f_w(x_i))^2}{2\sigma^2} \right\}. \end{aligned} \tag{1}$$

Maximizing expression (1) is equivalent to maximizing the logarithm of the expression (since log is monotone increasing), which is equivalent to *minimizing* the negative logarithm of the expression. Thus (1) is equal to

$$\begin{aligned} \underset{w \in \mathbb{R}^m}{\operatorname{argmax}} L(w) &= \underset{w \in \mathbb{R}^m}{\operatorname{argmin}} \sum_{i=1}^{n} \Big(\frac{(y_i - f_w(x_i))^2}{2\sigma^2} - \log \frac{1}{\sqrt{2\pi\sigma^2}} \Big) \\ &= \underset{w \in \mathbb{R}^m}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^{n} (y_i - f_w(x_i))^2 \\ &= \underset{w \in \mathbb{R}^m}{\operatorname{argmin}} E(w). \end{aligned}$$

Thus, the w we estimate using least squares regression is the same as the w given by maximum likelihood estimation. This derivation gives theoretical justification to the sum-of-squares error we used last lecture. Not only is this error function mathematically convenient, it is based on the statistical principle of maximum likelihood estimation. Keep in mind, however, that a fair number of assumptions were required in our analysis. As a result, least squares regression may not always be the most appropriate model to use.

Least squares regression is sensitive to outliers

Because we are using squared errors, outliers have a lot of sway in biasing our final set of weights. To make our model more robust to outliers, we now suppose

$$y_i = \begin{cases} f_{\hat{w}}(x_i) + \epsilon_i & \text{with probability } 0.99 \\ U(-a, a) & \text{with probability } 0.01, \end{cases}$$

where $U(-a, a)$ denotes a random variable that is uniformly distributed over the interval $(-a, a) \subseteq \mathbb{R}$. Then we have

$$P(y_i | x_i, \hat{w}) = 0.99 \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big\{ -\frac{(y_i - f_{\hat{w}}(x_i))^2}{2\sigma^2} \Big\} + 0.01 \cdot \frac{1}{2a}.$$

Unfortunately, maximizing the likelihood is intractable when the probability takes this form.

Mean and median

The sample mean of a sample $X_1, \ldots, X_n$ is given by $\frac{1}{n}\sum_{i=1}^{n} X_i$. You may have heard that the sample mean is sensitive to outliers. Consider, for example, how the mean changes when a sample $\{3, 4, 4, 5\}$ is updated to $\{3, 4, 4, 5, 14\}$.

The sample median can be computed by sorting the values in a sample and taking the middle one. In contrast to the sample mean, the sample median is an example of a summary statistic that is not sensitive to outliers. (Try computing the sample median for each of the samples above.)

An alternative (but equivalent) way to define the sample mean μ^* is

$$\mu^* = \operatorname*{argmin}_{\mu} \sum_{i=1}^{n} (x_i - \mu)^2.$$

Similarly, the sample median m^* can be written as

$$m^* = \operatorname*{argmin}_{m} \sum_{i=1}^{n} |x_i - m|$$

Next lecture we will look at an error function motivated by this formulation of the median, i.e.

$$E(w) = \sum_{i=1}^{n} |f_w(x_i) - y_i|$$

where

$$y_i = f_w(x_i) + \epsilon_i$$

and $\epsilon_i \sim$ Laplace(0,b) where the density of a Laplace(0,b) distribution is given by

$$\frac{1}{2b} \exp\Big\{ - \frac{|x - \mu|}{b} \Big\}.$$