CS142: Machine Learning Spring 2017

Lecture 8

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

Nonparametric Estimation

We continue our discussion of nonparametric estimation from last lecture. Recall that in nonparametric estimation, we make no assumption about the particular functional form of the underlying distribution from which data are sampled.

Histograms

Let's review some of the results we derived at the end of last lecture. If we have some region R about a point x in our sample space, the density at x can be approximated by

$$p(x) \approx \frac{K}{NV}, \tag{1}$$

where K is the number of examples (from some training data set) in R, N is the total number of examples, and V is the volume of the region R. Note that when our sample space is the real line, this formula dictates how to compute the heights of the bars in a histogram: regions become intervals and volumes become interval lengths. (See equation 2.241 in the book.)

We can apply the above result by first chopping up our sample space into equally sized disjoint "bins." If we let K_b be the number of example points in bin b, then our estimate becomes

$$p(x) = \frac{K_b}{NV},$$

where b is the bin containing x, N is the size of the training set, and V is the volume of a bin. The number of bins grows with h^D where h is the number of intervals in each dimension. This implies that the size of our training set needs to increase exponentially with the dimensionality of the sample space in order to avoid a situation in which most of the bins are empty.

Example: Suppose we have two classes $Y = \{1, 2\}$ which, *a priori*, we consider to be equally likely. Then for any $x \in X$ we have

$$p(y \mid x) \propto p(x \mid y)p(y) = p(x \mid y) \cdot \frac{1}{2}.$$

Given a training set $T = \{(x_1, y_1), \ldots, (x_N, y_N)\}$, we can create two histograms, one for each class. This gives us estimates of $p(x \mid y = 1)$ and $p(x \mid y = 2)$. We have a assumed a uniform prior, so the MAP estimate of the class y for a new point x is simply whichever class has a higher histogram bar for the bin that x belongs to.

Parzen windows

The histogram approach to nonparametric estimation inspires the more general, tunable *Parzen window method*. (In APMA 1740 and the textbook, this method is referred to as *kernel density estimation*.) This approach is defined by a kernel function $g(u)$ satisfying

$$g(u) \geq 0, \quad \int g(u)du = 1. \tag{2}$$

In other words, $g(u)$ satisfies the properties of a probability density function. One example of a kernel for a D-dimensional sample space is

$$g(u) = \begin{cases} 1 & |u_i| \leq \frac{1}{2} \text{ for all } i \in \{1, \ldots, D\} \\ 0 & \text{otherwise.} \end{cases}$$

超立方体

The kernel $g(u)$ is 1 whenever u falls within the D-dimensional unit hypercube centered about the origin. If we have a sample $\{x_1, \ldots, x_N\}$, then

$$K = \sum_{i=1}^{N} g\left(\frac{x - x_i}{h}\right)$$

is the number of sample points falling within the the hypercube with side length h centered at the point x. Substituting this result into equation (1) from the last section, we obtain the following estimate for the density at x,

$$p(x) = \frac{1}{Nh^D} \sum_{i=1}^{N} g\left(\frac{x - x_i}{h}\right) \tag{3}$$

where the volume of the hypercube h^D has replaced V.

Define the function

$$g_i(x) = \frac{1}{h^D} \cdot g\left(\frac{x - x_i}{h}\right)$$

for each $i \in \{1, \ldots, N\}$. Then

$$p(x) = \frac{1}{N} \sum_{i=1}^{N} g_i(x).$$

So $p(x)$ is a mixture of the densities $g_i(x)$ for $1 \leq i \leq N$.

Note that the right-hand side of equation (3) will always integrate to one, as long as the conditions in (2) are satisfied. In other words, the hypercube kernel is just one of many possible choices for $g(u)$. Equation (3) is general and holds for any kernel function.

K-nearest neighbors classification

A simple way to estimate the class y of a sample point x is to examine the K examples that are closest to x in the training data set. Note that we need an appropriate definition of distance in order to assess which points in the training set are close to x. Euclidean distance is a standard choice. After identifying the K closest training points, we take a majority vote; i.e. we assign x to the class that has the largest representation among these K training points.

Bayesian vs. Maximum Likelihood Estimation

Let $D = \{x_1, \ldots, x_N\}$ be a set of i.i.d random variables distributed according to a Bernoulli(μ) distribution. Let m be the number of 1's in the sample D and l be the number of 0's. The likelihood $L(\mu \mid D) = P(D \mid \mu)$ is

$$P(D \mid \mu) = \mu^m (1-\mu)^l.$$

In the last lecture, we solved for the MLE estimate of μ,

$$\mu_{\text{MLE}} = \frac{m}{m+l}.$$

Now consider the posterior probability $P(\mu \mid D)$. By Bayes' theorem, we can write this probability as

$$P(\mu \mid D) = \frac{P(D \mid \mu)P(\mu)}{P(D)} \propto P(D \mid \mu)P(\mu), \tag{4}$$

where $P(D \mid \mu)$ is the likelihood and $P(\mu)$ is the prior distribution on the parameter μ.

The beta distribution

A convenient prior for the parameter μ of a Bernoulli distribution is the beta distribution. A beta distribution is defined by two parameters, α and β, and has the following probability density function:

$$P(\mu \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\mu^{\alpha-1}(1-\mu)^{\beta-1}, \qquad \mu \in [0,1]$$

where Γ is the gamma function. The left fraction in the above expressions can be thought of as a constant to ensure that the density integrates to one. The interesting part of the density function is the stuff on the right, which should look familiar.

As an aside, note that if $\alpha = \beta = 1$, the corresponding beta distribution is a Uniform(0,1) distribution.

Plugging this distribution in as the prior on the right-hand side of (4), we have

$$\begin{aligned} P(\mu \mid D) &\propto \mu^m(1-\mu)^l\mu^{\alpha-1}(1-\mu)^{\beta-1} \\ &= \mu^{m+\alpha-1}(1-\mu)^{l+b-1} \\ &= P(D' \mid \mu). \end{aligned}$$

The last equation above is meant to convey the fact that the posterior distribution of μ can be regarded as (proportional to) the likelihood of a new data set D', which is constructed from the original data set D by adding $\alpha - 1$ ones and $\beta - 1$ zeros to the sample.

We use the posterior distribution to compute the MAP estimate for μ,

$$\mu_{\text{MAP}} = \operatorname*{argmax}_{\mu} P(\mu \mid D) = \frac{m + (\alpha-1)}{m + l + (\alpha-1) + (\beta-1)}.$$

So the MAP estimate for μ is equal to the maximum likelihood estimate using an augmented data set D'. Further, observe that the posterior distribution $P(\mu \mid D)$ is yet another beta distribution, i.e.

$$P(\mu \mid D) = \text{Beta}(\mu \mid m+\alpha,\, l+\beta).$$

Suppose we observe N tosses and we want to estimate the probability of the next toss turning out heads. Then we want to compute

$$P(x_{N+1} \mid x_1, \ldots, x_N).$$

This probability is

$$P(x \mid D) = \int_{\mu\in[0,1]} P(x \mid \mu)P(\mu \mid D)\, d\mu.$$

It follows that

$$P(x = 1 \mid D) = \frac{m+\alpha}{m+\alpha+l+\beta}.$$