CS142: Machine Learning Spring 2017

Lecture 7

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

Bayes optimal classifier with 0,1-loss

Last lecture we found that the Bayes optimal classifier $\hat{y} = f(x)$ is defined by

$$\begin{aligned}\hat{y} &= \underset{y}{\operatorname{argmax}}\, p(y|x) \\ &= \underset{y}{\operatorname{argmax}}\, p(x|y)p(y),\end{aligned}$$

where $p(y)$ is a discrete distribution and we assume that $p(x|y)$ takes the parametric form of $p_{\theta_y}(x)$. For example, in the fish classification example, $p_{\theta_y}(x)$ was assumed to be a normal distribution and θ_y could be either (μ_S, σ_S) or (μ_B, σ_B). Note that in this computation, we have assumed that we know the joint distribution $p(x, y)$.

Parametric distributions $p_\theta(x)$

Suppose we have a training dataset T of i.i.d. samples $X_1, X_2, \ldots, X_N$ from $p_\theta(x)$. We look for an estimator $\hat{\theta} = f(X_1, \ldots, X_N)$. Ideally, we would like the following properties for our estimator f.

1. We want f to be **unbiased**. In other words, if θ is the true value of the parameter we wish to estimate, then we would like

$$E[f(X_1, \ldots, X_N)] = \theta.$$

渐近地

2. If f is not unbiased, we may accept the weaker condition of f being **asymptotically unbiased**, i.e.

$$\lim_{N\to\infty} E[f(X_1, \ldots, X_n)] = \theta.$$

3. Finally, we would like for f to be **consistent**, i.e. for any $\varepsilon > 0$, we want

$$\lim_{N\to\infty} P(\|f(X_1, \ldots, X_n) - \theta\| > \varepsilon) = 0.$$

Maximum likelihood estimation gives an asymptotically unbiased and consistent estimator in most cases.

Example: Let $p_\theta(x) = \mathcal{N}(x, \mu, \sigma^2)$. Suppose we know that $\sigma^2 = 1$. Then the unknown parameter θ is the mean μ. Consider the (bad) estimator

$$f(X_1, \ldots, X_n) = X_1.$$

Then f is unbiased, since $E[f(X_1, \ldots, X_N)] = E[X_1] = \mu$, but f is not consistent.

Consider another estimator,

$$f(X_1, \ldots, X_N) = \frac{1}{N}\sum_{i=1}^{N} X_i,$$

i.e. the sample mean. You should verify that this estimator is both unbiased and consistent.

Bernoulli MLE

Let $X \sim \text{Bernoulli}(\mu)$ with $\mu \in [0, 1]$, meaning that $P(X = 1) = \mu$ and $P(X = 0) = 1 - \mu$. In this case, the parameter θ that we wish to estimate is μ. We can more concisely write the pmf of X as

$$p(x) = \mu^x (1-\mu)^{1-x}.$$

p(x=0) = 1 - u
p(x=1) = u

(You should check that for $x = 0$ and $x = 1$, this equation matches the two probability equations above.) Let $T = \{x_1, \ldots, x_n\}$ be a training set (assumed i.i.d.). The likelihood function L is

$$\begin{aligned} L(\mu|T) &= P(T|\mu) \\ &= \prod_{i=1}^{n} p(x_i|\mu) \\ &= \prod_{i=1}^{n} \mu^{x_i}(1-\mu)^{1-x_i}. \end{aligned}$$

$x_i = 0, 1$

We wish to compute the maximum likelihood estimator

$$\hat{\mu} = \operatorname*{argmax}_{\mu} P(T|\mu).$$

Denote the log likelihood as $l(\mu|T) = \log L(\mu|T)$, so

$$l(\mu|T) = \sum_{i=1}^{n} x_i \log \mu + (1 - x_i). \log(1 - \mu)$$

Since the logarithm is a monotonically increasing function, we have

$$\hat{\mu} = \underset{\mu}{\operatorname{argmax}}\ l(\mu|T).$$

Taking the derivative of this expression with respect to μ, we get

$$\frac{\partial l}{\partial \mu} = 0 \implies \sum_{i=1}^{n} \frac{x_i}{\mu} + \frac{1-x_i}{1-\mu} \cdot (-1) = 0.$$

If we let k be the number of ones in the training set and n be the total number of samples, the above equation can be rewritten as

$$\frac{k}{\mu} + \frac{n-k}{1-\mu} \cdot (-1) = 0.$$

Solving for μ, we find that

$$\hat{\mu}_{\text{MLE}} = \frac{k}{n}.$$

Discrete categorical distribution

Suppose X is a random variable that takes values in the set $\{1, \ldots, k\}$. We define the distribution of X to be

$$P_\theta(X = i) = \theta_i,$$

where $\theta_i \in \mathbb{R}^k$, $0 \leq \theta_i \leq 1$, and $\sum_{i=1}^{k} \theta_i = 1$. The likelihood of an i.i.d. training set $T = \{x_1, \ldots, x_n\}$ is then

$$P(T|\theta) = \prod_{i=1}^{n} \theta_{x_i}.$$

Exercise: Show that the maximum likelihood estimate of θ_i is

$$\hat{\theta}_i = \frac{c_i}{n},$$

where c_i is the number of samples that take on value i. (Hint: use Lagrange multipliers).

Multivariate normal distribution

We say X is distributed according to a **multivariate normal** if $X \in \mathbb{R}^D$, $\mu \in \mathbb{R}^D$, $\Sigma \in \mathbb{R}^{D \times D}$, and the density of X is given by

$$\boxed{\mathcal{N}(x, \mu, \Sigma) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\Big\{ -\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu) \Big\}.}$$

It can be shown that the maximum likelihood estimates of μ and Σ are

$$\boxed{\begin{aligned}\hat{\mu}_{\text{MLE}} &= \frac{1}{N}\sum_{i=1}^{N} x_i,\\ \hat{\Sigma}_{\text{MLE}} &= \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu}_{\text{MLE}})(x_i - \hat{\mu}_{\text{MLE}})^T,\end{aligned}}$$

where each $x_i \in \mathbb{R}^D$.

Nonparametric estimation

Up until now, we have focused exclusively on **parametric** estimation: the (univariate) normal distribution is defined by two parameters μ and σ, the Bernoulli distribution is defined by a single parameter μ, and the multivariate normal distribution is defined by $d + d^2$ parameters in the form of the mean vector μ and covariance matrix Σ. In all of these cases, the parameters tell you everything you need to know about the distribution. But sometimes it may not be appropriate to assume that your data are generated according to any particular parametric form. Why, for example, should it be true that fish lengths are normally distributed? Who's to say that nature behaves so nicely?

The idea of **nonparametric** estimation is to estimate a density without assuming any particular form of the density (i.e. without assuming the distribution can be uniquely defined by a set of parameters).

Suppose we have a training set $T = \{x_1, \ldots, x_N\}$ where each sample $x_i \in \mathbb{R}^D$ is distributed according to $p(x)$. Fix a point $x \in \mathbb{R}^D$ and consider a region R centered at x. For this region, we define the probability $P \doteq P(X \in R)$ to be

$$P = \int_R p(x)dx.$$

Suppose that out of our N samples from $p(x)$, K samples that fall within the region R. Each sample falls within R with probability P. So K follows a binomial distribution,

$$\text{Binom}(K \mid N, P) = \frac{N!}{K!(N-K)!}P^K(1-P)^{N-K}.$$

Thus the expectation of and variance of K/N are given by

$$\boxed{\text{E}[K/N] = P} \tag{1}$$

$$\boxed{\text{Var}[K/N] = \frac{P(1-P)}{N}} \tag{2}$$

If N is sufficiently large, then $K/N \approx P$. If $p(x)$ is approximately constant over R, then $P \approx V p(x)$, where V is the volume of R, so

$$p(x) \approx \frac{P}{V} \approx \frac{K}{NV}.$$

Histograms

We can partition $\mathbb{R}^D$ into hypercubes with side length h. Letting these hypercubes serve as the regions we mentioned above, we have

$$p(x) = \frac{K}{Nh^D}.$$

The number of bins of this histogram grows exponentially in D. We check that $p(x)$ "integrates" to 1. Indeed,

$$\int p(x)dx = \sum_{\text{bins } b} V \frac{K_b}{NV} = 1,$$

where K_b is the number of samples that fall in bin b.