HW01 STAT W4400

Ethan Grant uni: erg2145

September 19, 2016

1. (a)

$$y_{new} = f(x_{new}) = argmax P(x|y)p(y) = p(x|y)P(y)$$

Naive Bayes Assumes independence of all characteristics thus:

$$P(x|\mu) = \Pi_{j=1}^{5} p(x^j, \mu_k)P(C_k)$$

By subbing in we get:

$$f(x_{new}) = argmax_{k\epsilon 1,2,3} p(x^j, \mu)P(C)$$
$$argmax_{k\epsilon 1,2,3} \Pi_{j=1}^{5} g(x^j|u_k^j, 1)P(y = k)$$

(b) You estimate the parameters using maximum likelihood for each dimension separately.

Class conditional distribution parameter that must be estimated is μ_k^j where k is the class and j is the dimension. WE know from lecture that the MLE for μ is the average for all values of a class at a given dimension. Thus:

$$\hat{u_k^j} = \frac{1}{|C_k|} * \sigma x_i^j$$

The class prior is estimated based on frequency using MLE where n is the total number of observations

$$P(\hat{y = k}) = \frac{|C_j}{n}$$

(c) I do expect the naive Bayes classifier to preform well because one of the weakness of the naive bayes is the assumption regarding Independence of each dimension of a given point. There is Independence of dimensions with spherical Gaussian.

2. (a) Cookbook for MLE

   i. ensure that the data is i.i.d

   ii. Take the log of the formula to turn it into a sum as opposed to multiplication. You can take the log in this case without changing the function b/c the max doesn't change b/c log is monotonically increasing on R_+

   iii. take the derivative with respect to the variable that you want to find the estimator for

   iv. set the derivative equal to zero and solve for the estimator

   v. check that the second derivative is less than zero to ensure the point found is a maximum

   (b)

$$\begin{aligned}
&\sum \nabla_\mu(ln(\frac{v}{\mu})) + ln(x^{v-1}) - ln(\Theta(v)) + ln(e^{\frac{-vx}{\mu}})) \\
&= \sum \nabla_\mu(ln(\frac{v}{\mu})) + \nabla_\mu(ln(x^{v-1})) - \nabla_\mu(ln(\Theta(v))) + \nabla_\mu(\frac{-vx}{\mu}) \\
&= \sum v * (\frac{1}{\frac{v}{\mu}}) * \frac{-v}{\mu^2} + \frac{-vx}{\mu^2} \\
0 &= \sum \frac{-v}{\hat{\mu}} + \frac{v * x_i}{\hat{\mu^2}} \\
&= \sum \frac{-v}{\hat{\mu}} + \sum \frac{v * x_i}{\hat{\mu^2}} \\
&= \frac{-n * v}{\mu} + \sum \frac{v * x_i}{\mu^2} \\
\frac{n * v}{\mu} &= \sum \frac{v * x_i}{\mu^2} \\
n * v &= \frac{v}{\hat{\mu}} * \sum x_i \\
\hat{\mu} = \frac{\sum x_i}{n} &= \frac{1}{n} * \sum x_i
\end{aligned}$$

Now we must check the second derivative to ensure this is a max

$$\begin{aligned}
&\sum \frac{v}{\mu^2} - \frac{2 * v * x * \mu}{\mu^2 * \mu^2} \\
&= \sum \frac{v}{\mu^2} - \frac{2 * v * x}{\mu^3} \\
0 &= \frac{n * v}{\hat{\mu^2}} - \frac{2 * v}{\hat{\mu^3}} * \sum x_i \\
&= \frac{n * v}{\bar{x^2}} - \frac{2 * v * n}{\bar{x^2}} \\
&= -\frac{n * v}{\bar{x^2}}
\end{aligned}$$

since n and v are both positive this result is negative confirming that this is a max

(c)

$$0 = \sum [\nabla_v (ln(\frac{v}{\mu}) + ln(x^{v-1}) - ln(\tau(v) + ln(e^{\frac{-v*x}{\mu}}))]$$

$$= \sum \nabla_v (ln(\frac{v}{\mu})) + \nabla_v (ln(x^{v-1})) - \nabla_v (ln(\tau(v))) + \nabla_v (\frac{-v*x}{\mu})$$

$$= \sum ln(\frac{v}{\mu}) + \frac{1}{\frac{v}{\mu}} * \frac{1}{\mu} * v + \nabla_v ((v-1) * ln(x_i)) - \phi - \frac{x_i}{\mu}$$

$$= \sum ln(\frac{\hat{v}}{\mu}) - (\frac{x_i}{\mu}) - 1) - \phi(\hat{v}) + ln(x_i)$$

$$= \sum ln(\frac{x_i * \hat{v}}{\mu}) - (\frac{x_i}{\mu} - 1) - \phi(\hat{v})$$

3. By the definition of the classifier : $f_0(x) = argmax P(y|x)$ Pr(error) = risk under zero-one loss by definition

   Thus, one must minimize (based on hint 2):

$$R(f|x) = \sum_{y \epsilon k} L^{0-1}(y, f(x)) P(y|x)$$

$$= \sum_{y != k} L^{0-1}(f(x), k) P(k|x) + 0$$

$$= \sum_{y != k} P(k|x)$$

   We know that $\sum_{y \epsilon k} P(k|x) = 1$ because the x must be classified and if that probability were less than 1 it would imply there was some nonzero chance of it not being put in any of the classes which definitionally cannot occur. Thus since we are only excluding in the summariton when y=k

$$R(f|x) = \sum_{y != k} P(k|x) = 1 - P(f(x)|x) \quad (1)$$

   Since we know that by definition we have already maxxed P(f(x)—x) we also must minimized the risk since any other classifier will decrease P(f(x)—x) and cause the risk to increase.

4. (a) $R(f) = \sum_{y=1}^{k} \int L(y, f(x)) p(x, y) dx$
   $L(y, f(x))$ = the loss function $p(x, y)$ = the joint density function

   (b) Empirical risk is used because people don't know the true joint density function that is required to use the estimated risk function

(c) $lim_{n->\infty}|R(f) - \hat{R}_n(f) = 0$
This is true because across infinite independent draws will cause the estimation of $(\hat{R})_n$ to get increasingly close to the actual risk ratio (as described by the regression to the mean). Thus $(\hat{R})_n(f) = R(f)$

(d) $[0, 1]$ is the range because at worst it missess all its predictions and equals 1 or it gets all predicitions right and equals zero. R(f) is the probability of a mistaken prediction

(e) .5 because you would expect a random classifier to preform as well as randomly classifying which would get 50% of the predicitons correct

(f) The risk is smaller $E[R(f^2)] < E[R(f')]$ because it chooses better than at random according to the definition and thus has a lower probability of making an incorrect prediction. Since R(f) is the risk of an incorrect prediction that will be smaller for f^2