Rodgers (2017) A First Course in ML

5.2 Probabilistic Classifiers

Probabilistic = output is prob of a new object belonging to a particular class

train data in Matrix, vector = (X, t)

Probability for class c:

$P(T_{new} = c \mid x_{new}, X, t)$

5.2.1 ~~Naive Bayes~~ The Bayes classifier

5.2.2 logistic Regression

5.3.2 Support vector machines

Bayes Classifier

- takes name from equation it is based on
- Aim is to compute predictive probs of Classes

From Bayes Rule, expression for Predictive Prob:

$P(T_{new} = C \mid X_{new}, X, t) =$

Bayes Classifier

from Bayes Rules, Expression for pred prob:

$$P(T_{new} = c \mid X_{new}, X, t) = \frac{P(X_{new} \mid T_{new} = c, X, t) P(T_n = c \mid X, t)}{P(X_{new} \mid X, t)}$$

① $P(X_{new} \mid T_{new} = c, X, t)$

- the probability of observing the new feature vector X_{new} given that the new data point belongs to class c
  + given training set X & parameters t
- the likelihood of observing the new data points features if we assume it belongs to class C

② $P(T_{new} = c, X, t)$

- The Prior Probability that the new data point T_{new} belongs to class C
  + given training set X & parameters t

③ $P(X_{new} \mid X, t)$ ← marginal likelihood

- Probability of observing the new feature vector X_{new} given X training set & parameters t

the marginal liklihood can be expanded to a sum over all possible C class

$$p(X_{new} \mid X, t) \rightarrow$$

$$\sum_{c'=1}^{C} p(X_{new} \mid T_{new} = c', X, t) P(T_{new} = c' \mid X, t)$$

where c' = All possible classes

Why this expantion work? law of total prob

this says the prob of an event occuring (X_{new}) can be calculated by summing the probs occuring under each condition, i.e, the classes

Data is assumed to be complete & Mutually Exclusive, i.e. all classes known and data only belongs to one class

This Expantion is the product of the liklihood & prior prob for each class

How to define/identify these new eqs?

$p(X_{new} \mid T_{new} = c', X, t)$ = liklihood of X_n belong to Class C

$P(T_{new} = c \mid X, t)$ = prior prob of class C

note the full Expanded looks like:

$$= \frac{P(X_{new} \mid T_{new} = c, X, t)\, P(T_{new} = c \mid X, t)}{\sum_{c'=1}^{C} P(X_{new} \mid T_{new} = c', X, t)\, P(T_{new} = c' \mid X, t)}$$

This equation works by allowing us to calc the probability of a new data point belonging to a specific class by comparing the liklihood & prior for that class to the liklihood & prior for all classes

Same eqn on top & bottom w/ addition of Sum

How to calculate likelihood?

$$p(X_{new} \mid T_{new} = c, X, t)$$

This is a distribution specific to the c^{th} class ($T_{new} = c$) evaluated at X_{new}

To create a Bayes classifier, we need to

- Define C of these class conditional distributions
- it is common to use same dist for ea class
- But no strict ruling to do this
- Dist should be based on data being modelled & knowledge we have about data
- once dist chose, select parameters
- e.g. gaussian, mean & co-variance

How to calc prior probability?

$P(T_{new} = c \mid X, t)$

This is the probability that an object belongs to class c on just the training set, (X, t)

This piece of info allows us to give prior beliefs ~~in~~ about the class of a new data instance X_{new} before we see it

if $P(T_n = c, X, t)$ is really low then there exists a prior bias against X_{new} being c

we would only classify it as c if the the likelihood (update) is very high

the only restrictions on setting the probs is that they are positive & Σ to 1

two popular choices are:

① uniform prior: $P(T_{new}=c \mid X, t) = \frac{1}{C}$

② Class size prior: " $= \frac{N_c}{N}$

N = Train set N_c = Class num in train

Note ~~neither~~ dist does not need to be conditioned on X & t. ① uses neither ② uses t via C

?

Example Gaussian Class-Conditional

Each train object consists of a two-Dim attribute vector $x_n = [x_{n1}, x_{n2}]^T$ & an assoc label $t_n = \{1, 2, 3\}$

We can create a dataset from these 3 classes using distributions

This example uses gaussian as attributes are real valued

$$P(x_n \mid t_n = c, X, t) = N(\mu_c, \Sigma_c)$$

μ & Σ need to be chosen based on train pairs of class c, know X^c

This is now a machine learning class. we have X^c & we want to infer the parameters of the model: μ_c, Σ_c

find paras that max liklihood of obs X^c

Alternatively could use a Bayesian approach

Define a prior Density for these parameters, $p(\mu_c, \Sigma_c)$ to compute a posterior from Bayes rules

$$p(\mu_c, \Sigma_c \mid X^c) = \frac{p(X^c \mid \mu_c, \Sigma_c)\, p(\mu_c, \Sigma_c)}{p(X^c)}$$

Calc DIST of Paras given obs of X^c

then compute the likelihood of X_{new} by taking the expectation:

How likely DIF values of μ, Σ are after obs data

$$p(X_{new} \mid T_{new} = c, X, t) = E_p(\mu_c, \Sigma_c \mid X^c)$$

Calc average value by their distrib weights → $\{p(X_{new} \mid \mu_c, \Sigma_c)\}$ likelihood of X_n given μ, Σ

This equation calcs the likelihood of a new data point X_{new} belonging to class C by taking the Expectation of $p(X_n \mid \mu_c, \Sigma_c)$ with respect to the posterior distribution of μ_c, Σ_c

Expectation means calc the average of a function w/ respect to a distribution function

Max likihood estimates for mean (μ) & var (Σ) of a gaussian given N data points is obtained by Diff the log-likihood w/ respect to each parameter, setting to 0 & solving

1. likihood = Prob of obs data given μ & Σ
2. instead of work w/ likihood, work w/ log-likihood $\ln(L)$. log is monotonic so preserves location of the max. log is easier to work for summing probs
3. Differentiation - Derivative tells you how log-likihood changes as you vary parameter
4. Set to 0 & Solve = values that Max log-likihood

Notes:

- have function that tells you how well your data fits a gaussian dist for diff paras
- take derive to find paras that best fits the data

(MLE)

the Max likihood estamets for the mean & var become

$$\mu_c = \frac{1}{N_c} \sum_{n=1}^{N_c} x_n$$

$$\Sigma_c = \frac{1}{N_c} \sum_{n=1}^{N_c} (x_n - \mu_c)(x_n - \mu_c)^T$$

How to make predictions?

Once you have class-conditional distributions & the prior you can make predictions

Example: compute class probs for

$$x_{new} = [2, 0]^T$$

$$P(T_{new} = c \mid x_{new}, X, t) = \frac{P(x_{new} \mid T_{new} = c, X, t)\, P(T_{new} = c \mid X, t)}{\sum_{c'=1}^{C} " c' " \quad " c' "}$$

Various calcs needed:

c	likelihood $P(x_n \mid T_n = c, \mu_c, \Sigma_c)$	Prior $P(T_n = c \mid X, t)$	Product
1	0.198	1/3	0.0046
2	0.0061	1/3	0.0020
3	0.0002	1/3	0.0001
		use $\frac{N_c}{N}$	0.0067 this col summed the denominator

x_{new} [2, 0]

$P(T=1) = 0.6890$

$P(T=2) = 0.3024$

$P(T=3) = 0.0087$