**Classification – Discrete Case**

Let the *random input vector* represent *D* features, where each feature can be any of *K* possible values, be denoted by .

Let the *random output scalar*, with *C* possible classes, be denoted by .

The joint probability mass function, , which is generally unknown, assigns the probability that a certain instance *x* belongs to class *y*.

Given input *X*, we classify the instance to class with some function .

For simplicity, assume each feature is binary i.e. can take on either 0 or 1.

Assume comes from a known probability distribution but has unknown parameters, and . The former is a vector with elements, and the latter is a matrix.

Let denote a feature vector with element index , and let *c* denote a scalar for class.

Recall that for two events *A* and *B*.

Using the chain rule of conditional probability also explains the equality above.

We assume that the probability of an arbitrary instance *x*, based on just the parameters given previously and not on anything else i.e. values of *x* features, belonging to a specific class *c* has the probability .

For the first factor, conditional on the class *Y*, *X* only depends on .

For binary features, , and the probability of the jth feature being 1 (turned on) conditional on class follows a Bernoulli distribution.

For the second factor, very simply.

Putting both factors back together, and substituting instance *t* with to use the index *i*, see below.

Denote the entire dataset for training by .

The **likelihood function** is then given below.

The **log-likelihood function** is then derived from the above function.

Computing the maximum likelihood estimate is done by setting the first derivative of the above to zero (0) and solving for and for empirical estimates and , respectively.

(1) .

(2)

Maximum A Priori

**Derivation based on Factored Posterior**

Conjugate Priors

**Distributions**

**Beta Distribution**

There are two (2) shape parameters, *a* and *b*.

When , has a single mode.

The **beta function** is given by .

Note that the **gamma function** is given by

**Dirichlet-Multinomial Model**

The **K-simplex**, denoted by , is

**Dirichlet Distribution**

The dirichlet distribution is a generalization of beta distribution beyond binary random variables. There are *k* parameters, expressed collectively as a vector by .

and is a generalization of the Beta function to *K* dimensions.

Each parameter can be alternatively expressed by . As , the PDF peaks more.

The normalizing constant is .

For random sample , the statistic is a sufficient statistic if , meaning both come from the same probability distribution.