1 Derivation Approach 1

Given a set of features F where selected features are denoted as $f_i \in F$, we aim to select an optimal subset of features $F^* \subset F$ (where $|F_k^*| = k$ and $k < |F|$) relevant and not redundant to a given classifier and classification task. For computational efficiency, we will build F_k^* in a greedy manner by choosing the next optimal feature f_k^* given the previous set of optimal features $F_{k-1}^* = \left\{f_1^*, \ldots, f_{k-1}^*\right\}$ and recursively defining $F_k^* = F_{k-1}^* \cup \{f_k^*\}$ with $F_0^* = \emptyset$.

To begin the derivation, we provide a directed graphical model in Figure 1 to formalize the independence assumptions in probabilistic feature selection model for classification tasks. Shaded nodes represent observed variables while unshaded nodes are latent. The observed variables are the vector of attributes $\vec{x}^d$, the features f_i (where for $1 \leq i \leq k$, $f_i \in F$) and the actual label y^d. The y_i^d are binary random variables described by the conditional probability given $\vec{x^d}$ and f_i, where f_i are binary variables indicating whether the respective attribute x_i^d are relevant (1) or not (0) to the classification task.

The conditional probabilities table (CPTs) are as follows: $P(y_i^d|\vec{x}^d, f_i)$ represents the classification label prediction distribution given the data $\vec{x}^d$ and feature f_i.

We now formally define our initial objective:

$$f_k^* = \arg\max_{f_k} E_D[P\left(y^d|\vec{x}^d, F_{k-1}^*, f_k\right)]$$

Since jointly optimizing this objective is NP-hard, we take a greedy approach where we choose the best f_k^* assuming F_{k-1}^* is given. Then we can greedily optimize this objective as follows:

$$f_k^* = \arg\max_{f_k} E_D[P\left(y^d|\vec{x}^d, F_{k-1}^*, f_k\right)] \tag{1}$$

$$f_k^* = \arg\max_{f_k} \frac{1}{|D|} \sum_{d \in D} P(y^d|\vec{x}^d, F_{k-1}^*, f_k) \tag{2}$$

$$f_k^* = \arg\max_{f_k} \sum_{d \in D} \frac{P(y^d, \vec{x}^d, F_{k-1}^*, f_k)}{P(\vec{x}^d, F_{k-1}^*, f_k)} \tag{3}$$

$$f_k^* = \arg\max_{f_k} \sum_{d \in D} \sum_{\{y_i^d\}_{1 \leq i \leq k}} \frac{P(y^d, \vec{x}^d, F_{k-1}^*, f_k, \{y_i^d\}_{1 \leq i \leq k})}{P(\vec{x}^d, F_{k-1}^*, f_k)} \tag{4}$$

$$f_k^* = \arg\max_{f_k} \sum_{d \in D} \sum_{\{y_i^d\}_{1 \leq i \leq k}} \frac{P(y^d|\{y_i^d\}_{1 \leq i \leq k}) \prod_{i=1}^{k-1} \left(P(y_i^d|\vec{x}^d, f_i)\right) P(y_k^d|\vec{x}^d, f_k) P(\vec{x}^d) \prod_{i=1}^{k-1} \left(P(f_i)\right) P(f_k)}{P(\vec{x}^d, F_{k-1}^*, f_k)} \tag{5}$$

$$f_k^* = \arg\max_{f_k} \sum_{d \in D} \sum_{\{y_i^d\}_{1 \leq i \leq k}} P(y^d|\{y_i^d\}_{1 \leq i \leq k}) \prod_{i=1}^{k-1} \left(P(y_i^d|\vec{x}^d, f_i)\right) P(y_k^d|\vec{x}^d, f_k) \tag{6}$$

Here, we rewrote the expectation of a binary event as its probability, factorized the conditional probability in the joint probability divided by the marginal, marginalized over $\{y_i^d\}_{1 \leq i \leq k}$, factorized joint probability in conditional and prior following the graphical model and exploited d-separation to remove irrelevant conditions and cancel terms in the equation. Thus, we can optimize our initial objective aiming two goals of classification tasks: Precision and Recall. The following sections shows in detail how to develop these approaches.

1.1 Precision case

In order to select the subset of features providing high precision classifier we need the agreement of all features predictors y_i^d to do a precise prediction. Therefore, we need a conjunction operation between the predictors y_i^d. Considering a binary classification problem, when the actual label is true we need all of the predictors y_i^d equals to true. However, if the actual label were false, just one of the predictors y_i^d would have to be false. Thus the probability $P(y^d|\{y_i^d\}_{1 \leq i \leq k})$ can be expressed as follows:

$$P(y^d|\{y_i^d\}_{1 \leq i \leq k}) = I\left[y^d = \bigwedge_{i=1}^{k} y_i^d\right] = \begin{cases} 1; y^d = 1 \wedge \{y_i^d = 1\}_{1 \leq i \leq k} \\ 1; y^d = 0 \wedge \{y_i^d\}_{1 \leq i \leq k, \exists y_i^d \| y_i^d = 0} \\ 0; otherwise \end{cases} \tag{7}$$

According to these assumptions we can continue our derivation for achieve a high precision greedy feature selection algorithm. Here, we combined equations (**??**) and (**??**), separated each term according to the actual label value y^d and use the probability sum rule to rewire the second term as follows:

$$f_k^* = \arg\max_{f_k} \sum_{d \in D} I\left[y^d = \bigwedge_{i=1}^{k} y_i^d\right] \prod_{i=1}^{k-1} \left(P(y_i^d|\vec{x}^d, f_i)\right) P(y_k^d|\vec{x}^d, f_k) \tag{8}$$

$$f_k^* = \arg\max_{f_k} \sum_{d \in D} I[y^d = 1] \prod_{i=1}^{k-1} \left(P(y_i^d = 1|\vec{x}^d, f_i)\right) P(y_k^d = 1|\vec{x}^d, f_k) \tag{9}$$

$$+I[y^d = 0] \sum_{\substack{\{y_i^d\}_{1\leq i\leq k},\\ \exists y_i^d \| y_i^d = 0}} \prod_{i=1}^{k-1} \left(P(y_i^d|\vec{x}^d, f_i)\right) P(y_k^d|\vec{x}^d, f_k) \tag{10}$$

$$f_k^* = \arg\max_{f_k} \sum_{d \in D} I[y^d = 1] \prod_{i=1}^{k-1} \left(P(y_i^d = 1|\vec{x}^d, f_i)\right) P(y_k^d = 1|\vec{x}^d, f_k) \tag{11}$$

$$+I[y^d = 0] \left(1 - \prod_{i=1}^{k-1} \left(P(y_i^d = 1|\vec{x}^d, f_i)\right) P(y_k^d = 1|\vec{x}^d, f_k)\right) \tag{12}$$

From (**??**) we can intuitively describe how this equations is related with the precision metric defined in the classification evaluation confusion matrix. In the confusion matrix, Precision is calculated dividing the numbers of true positives by the number of all examples predicted as positive (in fact, true positives and false positives). When the actual label y^d is true the second term of (**??**) becomes 0 and the first term gives higher score for the feature f_k that provides a higher probability of predict true. Thus, we are stimulating the increasing of true positives in classification task. Meanwhile the actual label y^d is false, the first term of (**??**) becomes 0 and the second gives lower score to features f_k that have higher probability of label a given datum $\vec{x}^d$ as true. In this way we are penalizing features that generate false positive and reduce the amount of false positives. Thus, increasing true positives and reducing false positives we are directly increasing the precision metric.

1.2 Recall case

In order to select the subset of features providing higher recall classifier we need at least one feature predictors y_i^d predicting a given label to say that this datum is classified as this label. Therefore, we need a disjunction operation between the predictors y_i^d. Considering a binary classification problem, we need at least one predictors y_i^d equals to true when the actual label is true, in order to evaluate the disjunction operation as true. However, if the actual label were false, all of the predictors y_i^d would have to be false. Thus the probability $P(y^d|\{y_i^d\}_{1\leq i\leq k})$ can be expressed as follows:

$$P(y^d|\{y_i^d\}_{1\leq i\leq k}) = I\left[y^d = \bigvee_{i=1}^{k} y_i^d\right] = \begin{cases} 1; y^d = 1 \Rightarrow \{y_i^d\}_{1\leq i\leq k, \exists y_i^d \| y_i^d = 1} \\ 1; y^d = 0 \Rightarrow \{y_i^d = 0\}_{1\leq i\leq k} \\ 0; otherwise \end{cases} \tag{13}$$

According to these assumptions we can continue our derivation for achieve a high recall greedy feature selection algorithm. Here, we combined equations (**??**) and (**??**), separated each term according to the actual label value y^d and use the probability sum rule to rewrite the second term as follows:

$$f_k^* = \arg\max_{f_k} \sum_{d \in D} I\left[y^d = \bigvee_{i=1}^{k} y_i^d\right] \prod_{i=1}^{k-1} \left(P(y_i^d|\vec{x}^d, f_i)\right) P(y_k^d|\vec{x}^d, f_k) \tag{14}$$

$$f_k^* = \arg\max_{f_k} \sum_{d \in D} I[y^d = 0] \prod_{i=1}^{k-1} \left(P(y_i^d = 0|\vec{x}^d, f_i)\right) P(y_k^d = 0|\vec{x}^d, f_k) \tag{15}$$

$$+I[y^d = 1] \sum_{\substack{\{y_i^d\}_{1\leq i\leq k},\\ \exists y_i^d \| y_i^d = 1}} \prod_{i=1}^{k-1} \left(P(y_i^d|\vec{x}^d, f_i)\right) P(y_k^d|\vec{x}^d, f_k) \tag{16}$$

$$f_k^* = \arg\max_{f_k} \sum_{d \in D} I[y^d = 0] \prod_{i=1}^{k-1} \left(P(y_i^d = 0|\vec{x}^d, f_i)\right) P(y_k^d = 0|\vec{x}^d, f_k) \tag{17}$$

$$+I[y^d = 1] \left(1 - \prod_{i=1}^{k-1} \left(P(y_i^d = 0|\vec{x}^d, f_i)\right) P(y_k^d = 0|\vec{x}^d, f_k)\right) \tag{18}$$

From (**??**) we can intuitively describe how this equations is related with the recall metric defined in the classification evaluation confusion matrix. In the confusion matrix, Recall is calculated dividing the numbers of true positives by the number of all examples that actually are true (in fact, true positives and false negatives). When the actual label y^d is true, the first term of (**??**) becomes 0 and the second gives lower score to features f_k that have higher probability of label a given datum $\vec{x}^d$ as false. Thus, we stimulating features that generate true positive and consequently reduce the amount of false negatives since the sum of true positives and false negatives is a constant equal to the number of actual true examples in the data set. Thus, increasing true positives and reducing false negatives we are directly increasing the recall metric.

In addition, when the actual label y^d is false, the second term of (**??**) becomes 0 and the first gives higher score to features f_k that have higher probability of label a given datum $\vec{x}^d$ as false. Thus, we are stimulating the increasing of true negatives and consequently the decreasing of false positives which are not related to recall. However, it could be seen as a surrogate objective that is improve the accuracy metric in the classification task given accuracy is calculated dividing true positives and true negatives by the total number of data points.

2 Derivation Approach 2

$$f_k^* = \arg\max_{f_k} log\left(P(D|F_{k-1}^*, f_k)\right)$$

$$f_k^* = \arg\max_{f_k} log\left(\prod_{d\in D} P(y^d|\vec{x}^d, F_{k-1}^*, f_k)\right)$$

$$f_k^* = \arg\max_{f_k} \sum_{d\in D} log\left(\frac{P(y^d, \vec{x}^d, F_{k-1}^*, f_k)}{P(\vec{x}^d, F_{k-1}^*, f_k)}\right)$$

$$f_k^* = \arg\max_{f_k} \sum_{d\in D} log\left(\sum_{\{y_i^d\}_{1\leq i\leq k}} \frac{P(y^d, \vec{x}^d, F_{k-1}^*, f_k, \{y_i^d\}_{1\leq i\leq k})}{P(\vec{x}^d, F_{k-1}^*, f_k)}\right)$$

$$f_k^* = \arg\max_{f_k} \sum_{d\in D} log\left(\sum_{\{y_i^d\}_{1\leq i\leq k}} \frac{P(y^d|\{y_i^d\}_{1\leq i\leq k}) \prod_{i=1}^{k-1}\left(P(y_i^d|\vec{x}^d, f_i)\right) P(y_k^d|\vec{x}^d, f_k) P(\vec{x}^d) \prod_{i=1}^{k-1}\left(P(f_i)\right) P(f_k)}{P(\vec{x}^d, F_{k-1}^*, f_k)}\right)$$

$$f_k^* = \arg\max_{f_k} \sum_{d\in D} log\left(\sum_{\{y_i^d\}_{1\leq i\leq k}} P(y^d|\{y_i^d\}_{1\leq i\leq k}) \prod_{i=1}^{k-1}\left(P(y_i^d|\vec{x}^d, f_i)\right) P(y_k^d|\vec{x}^d, f_k)\right)$$

2.1 Precision case

$$P(y^d|\{y_i^d\}_{1\leq i\leq k}) = I\left[y^d = \bigwedge_{i=1}^{k} y_i^d\right] = \begin{cases} 1; y^d = 1 \wedge \{y_i^d = 1\}_{1\leq i\leq k} \\ 1; y^d = 0 \wedge \{y_i^d\}_{1\leq i\leq k, \exists y_i^d \| y_i^d = 0} \\ 0; otherwise \end{cases} \tag{19}$$

$$f_k^* = \arg\max_{f_k} \sum_{d\in D} log \left(I\left[y^d = \bigwedge_{i=1}^{k} y_i^d\right] \prod_{i=1}^{k-1} \big(P(y_i^d|\vec{x}^d, f_i)\big) P(y_k^d|\vec{x}^d, f_k) \right)$$

$$f_k^* = \arg\max_{f_k} \sum_{d\in D} log \left(I[y^d = 1] \prod_{i=1}^{k-1} \big(P(y_i^d = 1|\vec{x}^d, f_i)\big) P(y_k^d = 1|\vec{x}^d, f_k) + I[y^d = 0] \sum_{\substack{\{y_i^d\}_{1\le i\le k},\\ \exists y_i^d \| y_i^d = 0}} \prod_{i=1}^{k-1} \big(P(y_i^d|\vec{x}^d, f_i)\big) P(y_k^d|\vec{x}^d, f_k) \right)$$

$$f_k^* = \arg\max_{f_k} \sum_{d\in D} log \left(I[y^d = 1] \prod_{i=1}^{k-1} \big(P(y_i^d = 1|\vec{x}^d, f_i)\big) P(y_k^d = 1|\vec{x}^d, f_k) + I[y^d = 0](1 - \prod_{i=1}^{k-1} \big(P(y_i^d = 1|\vec{x}^d, f_i)\big) P(y_k^d = 1|\vec{x}^d, f_k)) \right)$$

$$f_k^* = \arg\max_{f_k} \sum_{d\in D} I[y^d = 1] log \left(\prod_{i=1}^{k-1} \big(P(y_i^d = 1|\vec{x}^d, f_i)\big) P(y_k^d = 1|\vec{x}^d, f_k) \right)$$

$$+ I[y^d = 0] log \left(1 - \prod_{i=1}^{k-1} \big(P(y_i^d = 1|\vec{x}^d, f_i)\big) P(y_k^d = 1|\vec{x}^d, f_k) \right)$$

$$f_k^* = \arg\max_{f_k} \sum_{d\in D} I[y^d = 1] \sum_{i=1}^{k-1} log \big(P(y_i^d = 1|\vec{x}^d, f_i)\big) + log \big(P(y_k^d = 1|\vec{x}^d, f_k)\big)$$

$$+ I[y^d = 0] log \left(1 - \prod_{i=1}^{k-1} \big(P(y_i^d = 1|\vec{x}^d, f_i)\big) P(y_k^d = 1|\vec{x}^d, f_k) \right)$$

2.2 Recall case

$$P(y^d|\{y_i^d\}_{1\le i\le k}) = I\left[y^d = \bigvee_{i=1}^{k} y_i^d\right] = \begin{cases} 1; y^d = 1 \Rightarrow \{y_i^d\}_{1\le i\le k, \exists y_i^d \| y_i^d = 1} \\ 1; y^d = 0 \Rightarrow \{y_i^d = 0\}_{1\le i\le k} \\ 0; otherwise \end{cases} \tag{20}$$

$$f_k^* = \arg\max_{f_k} \sum_{d\in D} log \left(I\left[y^d = \bigvee_{i=1}^{k} y_i^d\right] \prod_{i=1}^{k-1} \big(P(y_i^d|\vec{x}^d, f_i)\big) P(y_k^d|\vec{x}^d, f_k) \right)$$

$$f_k^* = \arg\max_{f_k} \sum_{d\in D} log \left(I[y^d = 0] \prod_{i=1}^{k-1} \big(P(y_i^d = 0|\vec{x}^d, f_i)\big) P(y_k^d = 0|\vec{x}^d, f_k) + I[y^d = 1] \sum_{\substack{\{y_i^d\}_{1\le i\le k},\\ \exists y_i^d \| y_i^d = 1}} \prod_{i=1}^{k-1} \big(P(y_i^d|\vec{x}^d, f_i)\big) P(y_k^d|\vec{x}^d, f_k) \right)$$

$$f_k^* = \arg\max_{f_k} \sum_{d\in D} log \left(I[y^d = 0] \prod_{i=1}^{k-1} \big(P(y_i^d = 0|\vec{x}^d, f_i)\big) P(y_k^d = 0|\vec{x}^d, f_k) + I[y^d = 1] \left(1 - \prod_{i=1}^{k-1} \big(P(y_i^d = 0|\vec{x}^d, f_i)\big) P(y_k^d = 0|\vec{x}^d, f_k) \right) \right)$$

$$f_k^* = \arg\max_{f_k} \sum_{d\in D} I[y^d = 0] log \left(\prod_{i=1}^{k-1} \big(P(y_i^d = 0|\vec{x}^d, f_i)\big) P(y_k^d = 0|\vec{x}^d, f_k) \right)$$

$$+ I[y^d = 1] log \left(1 - \prod_{i=1}^{k-1} \big(P(y_i^d = 0|\vec{x}^d, f_i)\big) P(y_k^d = 0|\vec{x}^d, f_k) \right)$$

$$f_k^* = \arg\max_{f_k} \sum_{d\in D} I[y^d = 0] \sum_{i=1}^{k-1} log \big(P(y_i^d = 0|\vec{x}^d, f_i)\big) + log \big(P(y_k^d = 0|\vec{x}^d, f_k)\big)$$

$$+ I[y^d = 1] log \left(1 - \prod_{i=1}^{k-1} \big(P(y_i^d = 0|\vec{x}^d, f_i)\big) P(y_k^d = 0|\vec{x}^d, f_k) \right)$$