Missing Values

1 A modification of Naive Bayes to deal with missing values

- Training

  When we fit $P(x_i|y)$ for feature x_i, we can just use all available values and ignore missing values.

- Testing[1]

  If a test data point has some missing features, say x_1, we can marginalizing it out.

$$\begin{aligned} P(y|x_2,...,x_d) &\propto P(y)P(x_2,...,x_d|y) \\ &= P(y)\sum_{x_1} P(x_1,x_2,...,x_d|y) \\ &= P(y)\sum_{x_1}\prod_i P(x_i|y) \\ &= P(y)(\sum_{x_1} P(x_1|y))\prod_{i=2,3,...,d} P(x_i|y) \\ &= P(y)\prod_{i=2,3,...,d} P(x_i|y) \end{aligned}$$

  So the classification rule essentially ignores the missing feature x_1, and uses other available feature values.

2 Missing values and Decision Trees and Stumps

There are several options:

- make every split (value $x^j?\theta$ threshold) have three branches: (bigger, $x^j > \theta?$) , (smaller, $x^j \leq \theta?$), and (missing x^j). The missing branch acts like any other branch, computing variance reduction, further splits etc.
- probabilistic splits: for split split (value $x^j?\theta$ threshold), first count how many non-missing points go directed to each of the two branches bigger-than and smaller-than; transform these counts into a distribution (bigger

p, smaller $1-p$). For the missing-value points, apply this distribution to obtain a probabilistic split: such datapoints will follow with probability p the bigger branch and with probability $1-p$ the smaller branch.

3 Missing values for linear models

Linear models include al model based at some step on linear combinations of values: regression, Logistic Regression, Perceptrons, Neural Network. For these models, missing values can be substituted with default values for the feature (typically mean or median)

References

[1] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[2] Maytal Saar-Tsechansky and Foster Provost. Handling missing values when applying classification models. *Journal of Machine Learning Research*, 2007.