CHEAT SHEET

Naive Bayes Classifier

Algorithm Name	Naive Bayes Classifier
Description	If our goal is to find the distribution for the label, namely $P(y\|\mathbf{x})$, we can use the naive Bayes classifier on this distribution to find the label of test points. To simplify the procedure for finding $P(y\|\mathbf{x})$, we assume that each feature is independent of the others given the label. Naive Bayes decomposes a d-dimensional probability estimation problem into d one-dimensional probability estimation problems, because MLE gets exponentially harder as d increases. We derive the naive Bayes classifier as follows: $h(\mathbf{x}) = argmax_y P(y\|\mathbf{x})$ $= argmax_y \frac{P(\mathbf{x}\|y)P(y)}{P(\mathbf{x})}$ (Bayes' rule) $= argmax_y\ P(\mathbf{x}\|y)P(y)$ ($P(\mathbf{x})$ does not depend on y) $= argmax_y \prod_{\alpha=1}^{d} P(x_\alpha\|y)P(y)$ (by the naive Bayes assumption) $= argmax_y \sum_{\alpha=1}^{d} log(P(x_\alpha\|y)) + log(P(y))$ (as log is a monotonic function)
Applicability	Classification problems where features can be assumed independent
Assumptions	Given the label, features are independent of one another.
Underlying Mathematical Principles	We assume the $\mathbf{x}_\alpha\|y$ follows some distribution (e.g., categorical distribution) and use MLE to learn the distribution from the data.
Additional Details	Optional +1 smoothing (Laplace smoothing)
Example	Email spam classification; features are words that appear in the emails, labels are spam/not spam.