Statistical learning approaches

Imagine that you need to design a spam-filtering algorithm starting from this initial (over-simplistic) classification based on two parameters:

Parameter	Spam emails (X_1)	Regular emails (X2)
p_1 - Contains > 5 blacklisted words	80	20
p_2 - Message length < 20 characters	75	25

We have collected 200 email messages (**X**) (for simplicity, we consider $\mathbf{p_1}$ and $\mathbf{p_2}$ mutually exclusive) and we need to find a couple of probabilistic hypotheses (expressed in terms of $\mathbf{p_1}$ and $\mathbf{p_2}$), to determine:

$$P(spam|h_{p1}, h_{p2})$$

We also assume the conditional independence of both terms (it means that $\mathbf{h_{p1}}$ and $\mathbf{h_{p2}}$ contribute conjunctly to spam in the same way as they were alone).

For example, we could think about rules (hypotheses) like: "If there are more than five blacklisted words" or "If the message is less than 20 characters in length" then "the probability of spam is high" (for example, greater than 50 percent). However, without assigning probabilities, it's difficult to generalize when the dataset changes (like...

❮ Previous Section (/book/big_data_and_business_intelligence/9781785889622/2/ch02lvl1s

Next Section ❯ (/book/big_data_and_business_intelligence/9781785889622/2/ch02lvl1sec1