Lo q' queda pendiente es calcular estimadores $P(X_i/S)$ y $P(X_i/\neg S)$.

Si tuviéramos una cantidad razonable de mails de entrenamiento, es natural estimar, por ejemplo $P(X_i/S)$ como el cociente de mails spam de entrenamiento q' contienen la palabra w_i sobre el total de mails q' dicen spam:

$$P(X_i/S) \sim \frac{\#\text{mails spam con } w_i}{\#\text{mails totales con spam}}$$

Problema: esta forma de estimar $P(X_i/S)$ tiene un problema. Supongamos q' la palabra "data" ocurre sólo en mails q' no son spam en nuestro conjunto de entrenamiento. Es decir: $P(\text{"data"}/S) = 0$

Entonces el clasificador asignará probabilidad 0 a cualquier mail q' contenga la palabra "data", aunque el texto sea

"data on cheap viagra and authentic rolex watches"

Solución: la estimación se hace utilizando una constante K (K tiene q' ser chico, ej: = 1) llamada pseudo-cuenta:

$$P(X_i/S) \sim \frac{(K + \#\text{mails spam con } w_i)}{(2K + \#\text{mails con spam})}$$

↓

se asigna un peso menor a las 'observaciones' ficticias.