Ch2.

2.1 Considering Stratified Sampling. there will be 150 positive samples and 150 negative samples for test set.

$\sim C_{500}^{150} \cdot C_{500}^{150}$

2.2

10-fold cross validation: (considering stratified sampling), the error rate will be 50%

Leave-One-Out : Definitely 100% wrong on test set

2.3.

$F_1 = \frac{2P \cdot R}{P+R}$ $\quad BEP: R = P$

When $P=R \Rightarrow F_1 = P = BEP$.

so at $BEP(A)$ & $BEP(B)$. if $F_1(A) > F_1(B)$, then $BEP(A) > BEP(B)$

2.4

$TPR = \frac{TP}{TP+FN}$ $\quad$ "=" $\quad P = \frac{TP}{P(T+F)}$

$FPR = \frac{FP}{FP+TN}$ $\quad R = \frac{TP}{TP+PN}$

2.5. 画图证明

2.6. (1)错误率说ROC曲线都是模型评估的方式.

(2)错误率评估的是模型在测试集上的效果, 而ROC曲线评估的是模型结果的分布情况.

(3)理想来说, 一个好模型错误率应可能接近于0. AUC应尽可能接近于1.

ROC每点对应一个错误率.

2.7. easily effected by outlier points (samples)

2.8.

	Pros	Cons
min-max	Simple.	if add a super big / small value, recalculate all from the beginning.
Z-score	~ normal distribution especially useful in some places	a lot of calculation (variance)

2.9.

1. Build Null Hypothesis,
2. Acquire χ^2 value from data.
3. compared with value in the table, decide whether or not to accept the null hypothesis.