یک.

a) Von Neumann entropy: $E = -\Sigma p_i \log_2 p_i$
In this case, $p_+ = 0.4,\ p_- = 0.6 \Rightarrow E = 0.673$
Conditioning on A, the entropy for A=F becomes 0, while the entropy for A=T becomes 0.683.
Normalizing the new entropies, we have $E_{splitA} = 0.7 * 0.683 = 0.478$
Conditioning on B, the entropy for B=T becomes 0.56, while the entropy for B=F becomes 0.45.
Normalizing on the new entropies, we have $E_{splitB} = 0.4 * 0.56 + 0.6 * 0.45 = 0.494$.
Less entropy is remained by conditioning on A, so A should be chosen.

b) Gini Impurity: $G = p_i * (1 - p_i)$, so the initial G is 0.48.
Splitting on A, the G for A=F becomes 0 while it's 0.49 for A=T. Normalizing on weights, we have $G_{splitA} = 0.343$
Splitting on B, the G for B=T becomes 0.375 while it's 0.278 for B=F. Normalizing on weights, $G_{splitB} = 0.4 * 0.375 + 0.6 * 0.278 = 0.3168$
The lower impurity after split, the better. So, B should be chosen.

دو.

a) P(+) = 0.5, P(-) = 0.5, Knowing $P(x|y) = \frac{P(x \cap y)}{P(y)}$, assuming P(X) = P(X=1), we have:
$P(A|+) = \frac{0.3}{0.5} = 0.6,\ P(B|+) = \frac{0.1}{0.5} = 0.2,\ P(C|+) = \frac{0.4}{0.5} = 0.8$
$P(A|-) = \frac{0.2}{0.5} = 0.4, P(B|-) = \frac{0.2}{0.5} = 0.4,\ P(C|-) = \frac{0.5}{0.5} = 1$

b) P(A'|+) * P(B|+) * P(C'|+) = 0.4 * 0.2 * 0.2 = 0.016
P(A'|-) * P(B|-) * P(C'|-) = 0.6 * 0.4 * 0 = 0 (!) (Noting that P(+) = P(-))
This means that it the label should be +, but since while performing Naïve Bayes classification, we usually want to avoid exact 0 probabilities, we could augment α additional instances of each possibility. Choosing $\alpha = 1$, our new probabilities become:

$P(A|+) = \frac{4}{8} = 0.5, P(B|+) = \frac{2}{8} = 0.25, P(C|+) = \frac{5}{8} = 0.625$
$P(A|-) = \frac{3}{8},\ P(B|+) = \frac{3}{8},\ P(C|-) = \frac{6}{8}$

P(A'|+) * P(B|+) * P(C'|+) = 0.5 * 0.25 * 0.375 = 0.0469
$P(A'|-) * P(B|-) * P(C'|-) = \frac{5}{8} * \frac{3}{8} * \frac{2}{8} = 0.05$

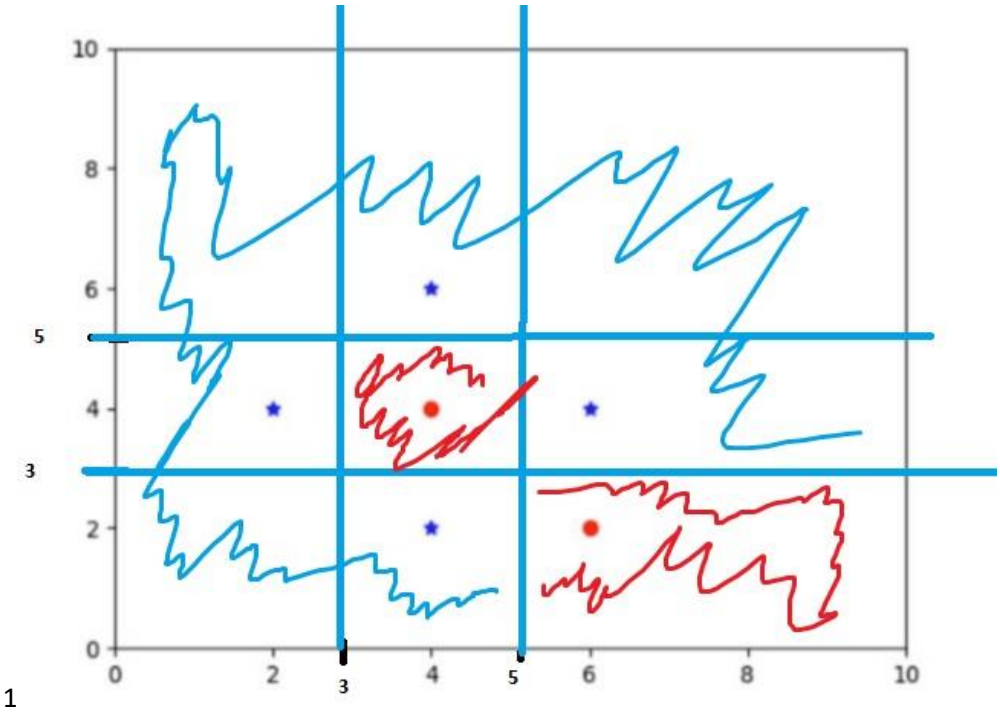
So, if we treat 0 probabilities as absolute 0, we should get a + label, but if we assume that have a value that's just currently missing (with an alpha of 1), we'll get a – label.

سه.

در یک setting واقعی، در مواقعی که تاثیرات این classification حاد هستند، Accuracy معیار خوبی نیست، مثلا فرض کنید که خروجی ۱ را معادل انجام عمل جراحی قلب باز و خروجی ۰ را معادل تجویز آسپرین در نظر بگیریم. در این جور مواقع False Positive ها بسیار مشکل‌زا می‌شوند و صرف Accuracy پاسخ‌گو نخواهد بود. یا مثلا در وضعیت‌هایی که متغیر مورد پیش‌بینی ما در عمل طیف پیوسته‌ای داشته‌باشد، صرف Accuracy می‌تواند تعداد زیادی از موارد لب مرزی را نادیده بگیرد. حتی در مواردی که Error Rate بالایی

داریم نباید به Accuracy بسنده کرد. مثلا در مدلی که احتمال دو خروجی آن (0.4, 0.6) باشند، اگر مدل ما همیشه حالت 0.6 را به عنوان خروجی در نظر بگیرد، Accuracy صد درصد داریم ولی در عمل مدل ما پیش‌بینی خاصی انجام نمی‌دهد.

چهار.



الف) چون در نقاط (5,5) ,(5,3) ,(3,3) ,(3,5) فاصله‌ی نقطه‌ی قرمز در 4,4 با تقاط آبی نزدیک آن یکسان می‌شوند، این ۴ نقطه را به عنوان مرزهای نقطه‌ی وسط در نظر می‌گیریم و آن ها را به صورتی که فاصله اقلیدسی حفظ شوند امتداد می‌دهیم. برای نقطه‌ی قرمز (6,2) نیز به همین ترتیب.

ب) نزدیک‌ترین نقاط به آن نقاط آبی در (4,6) ,(6,4) هستند که هر کدام فاصله‌ای برابر با جذر ۲۰ با آن دارند.

پ) بله، می‌شود. با میانگین وزن‌دار گرفتن از مقدار ویژگی‌های k همسایه‌ی نزدیک آن، مقدار آن نقطه برای حل مساله‌ی رگرسیون پیدا می‌شود.

ت) بستگی به ساختار مساله، شناخت ما از فیچرهای مهم، مقدار نویز موجود در دیتاست و تاثیر میزان K روی پرفرمنس مدل ما دارد. برای حل این مساله می‌توان مدل را چندبار با k های مختلف اجرا کرد و آنی که مقدار error rate کمتری دارد را به عنوان k اپتیمال انتخاب کرد.

ث) بستگی به ابعاد دیتا دارد، اگر تعداد بعدها پایین باشد می‌توان با استفاده از ساختار داده‌های خاصی استفاده از آن را ممکن کرد، وگرنه چون باید روی تعداد زیادی دیتا iteration انجام دهد و فاصله را با تمامی آن‌ها حساب کند، توصیه نمی‌شود. (مگر این‌که بتوان با استفاده از ساختارداده‌هایی همچون binary search tree تعداد datapointهای سرچ شده را تا حد خوبی کاهش داد.)

ج) پیچیدگی زمانی آن در هنگام training برای هر داده O(1) است چون پردازش خاصی روی دیتا انجام نمی‌دهد و صرفا آن را ذخیره می‌کند. در زمان تست اما اضافه کردن هر دیتای تست حدود O(n) زمان می‌برد.

چ) هنگامی که ابعاد مساله بالا باشد به جای فاصله اقلیدسی از فاصله منهتن استفاده می‌شود که به نسبت دقت کم‌تری دارد اما محاسبه‌ی آن بسیار سریع‌تر انجام می‌شود.

پنج.

یک شیوه‌ی آمار و احتمالاتی برای سنجش دقت یک مدل در هنگامی که تعداد نمونه‌ها محدود باشند استفاده می‌شود.
دسته‌های کلی آن موارد زیر هستند:
Exhaustive که تمامی حالات را بررسی می‌کند، Non-exhaustive که تنها بخشی از حالات را بررسی می‌کند (با استفاده از سمپلینگ) و Nested که تقسیم‌بندی‌های تو در تو انجام می‌دهد و زمانی به کار می‌آید که نیاز داریم hyperparameterها را نیز تنظیم کنیم.

با افزایش k، اندازه‌ی training data زیاد شده و test data کم می‌شود، در نتیجه بایاس کم‌تری به علت تنوع بیش‌تر داده آموزشی خواهیم داشت که افزایش واریانس را نتیجه می‌دهد. پیچیدگی زمانی هم زیاد می‌شود، چراکه فرآیند cross-validation باید k دفعه انجام شود.