CS109 Lecture 21 Worksheet: Comparing Classifiers

Naive Bayes, Train/Test Splits, Calibration, Precision/Recall, and Fairness

Summer 2026

Lecture 21 Worksheet: Comparing Classifiers

Topics: other models for classification (brute-force Bayes, Naive Bayes, decision trees), the train/test paradigm and overfitting, baselines, calibration, precision and recall, and the two philosophies of algorithmic fairness.

How to use this sheet. Work in small groups. A TA and the instructor will circulate to help. We will go over selected solutions on the board. The problems are short — you do not need to finish every one. Keep fractions exact where you can.

Useful facts. $\sigma(z) = \dfrac{1}{1+e^{-z}}$, $\sigma(0) = 0.5$, $\sigma(1) = 0.7311$, $\sigma(1.5) = 0.8176$, $\sigma(2) = 0.8808$. $H(p) = -p\log_2 p - (1-p)\log_2(1-p)$.

Problem 1 (Review of Lecture 20): Logistic Regression Refresher

A logistic regression with two binary features has learned $\theta = [\theta_0, \theta_1, \theta_2] = [-1,\ 3,\ -1]$ (recall $x_0 = 1$).

(a) Compute $P(Y = 1 \mid X = x)$ for the datapoint $(x_1, x_2) = (1, 1)$.

(b) The true label for that datapoint is $y = 1$. Compute the gradient contribution $\frac{\partial LL}{\partial \theta_1}$ from this single example. Would a gradient ascent step increase or decrease θ_1?

(c) In one sentence, what does the negative sign on θ_2 tell you about feature 2?

Problem 2: Brute Force Bayes and Why We Need an Assumption

A "brute force Bayes" classifier predicts $\hat{y} = \arg\max_y P(Y = y \mid X = x)$ by storing a full conditional probability table: for every possible assignment of the features and every class label, one probability. Assume all m features are binary and Y is binary.

(a) With $m = 2$ features, how many probabilities must the model learn? Sketch the table.

(b) Write the number of parameters as a function of m. What is the big-O?

(c) For the heart dataset, $m = 22$. For the ancestry dataset, $m = 100$. How many parameters would brute force Bayes need in each case? (You may leave the second as a power of 2, but compare it to the roughly 10^{80} atoms in the observable universe.)

(d) The problem is not really *storage* — it is *data*. Explain in one or two sentences why a table with 2^{100} entries cannot be estimated from 467 training datapoints, no matter how much memory you have.

(e) State the Naive Bayes assumption, and explain how it reduces the parameter count. What is the new big-O?

Problem 3: Naive Bayes With a Prior

(problem from the 2022 Fall Final)

A Naive Bayes classifier is trained on a dataset with 100 examples, 30 of which are labeled positive and 70 of which are labeled negative. Instead of using a Laplace prior, you use a Beta($a = 3, b = 4$) prior.

(a) What is your estimate for the probability $Y = 1$?

(b) The MLE estimate would have been 0.30. Your answer should be slightly different. Explain the direction of the shift in terms of what the prior's pseudo-counts are "pretending" to have observed.

(c) Suppose one of the 22 binary features never once takes the value 1 among the 30 positive examples. What would the MLE conditional probability $P(X_j = 1 \mid Y = 1)$ be, and what disaster does that cause at prediction time? How does a prior fix it?

Problem 4: Train, Test, and Overfitting

In lecture we compared several models on the same dataset:

Model	Train Accuracy	Test Accuracy
Baseline (always predict 1)	0.6031	0.5887
Naive Bayes	0.7909	0.8067
Logistic Regression	0.8169	0.8307
Decision Tree	0.8514	0.8307
Random Forest	0.8726	0.8500
Gradient Boosting	0.8611	0.8440

(a) A classmate reports that their model gets 99.8% accuracy. You ask one follow-up question. What is it, and why?

(b) Which model in the table shows the clearest sign of overfitting? Cite the specific numbers that make you say so.

(c) The baseline gets 58.87% on test *without learning anything.* Why is it important to report a baseline alongside every model?

(d) Random Forest beats Logistic Regression by about 2 percentage points on test. Before declaring a winner, what would you want to know about how that number was computed? (Think back to the bootstrapping lecture.)

(e) Logistic Regression and Naive Bayes both draw *linear* decision boundaries. Sketch a 2-D dataset that no line can separate. Despite this, both models often work well in practice — why?

Problem 5: Is My Model Calibrated?

Your heart-disease classifier is deployed and doctors want to use the output *probability*, not just the class. You bucket the 500 test patients by the model's output probability and count true labels:

Group	Avg. output prob.	Count($Y = 0$)	Count($Y = 1$)
Very low	0.1	92	8

Group	Avg. output prob.	Count($Y=0$)	Count($Y=1$)
Low	0.3	70	30
Borderline	0.5	48	52
High	0.7	25	75
Very high	0.9	30	70

(a) For each group, compute the observed fraction of patients with $Y = 1$. Which groups are well calibrated?

(b) What is the model's *accuracy* on the 100 patients in the "very high" group, if we threshold at 0.5?

(c) Assuming the model were perfectly calibrated at 0.9, what is the probability of observing exactly 70 patients with $Y = 1$ and 30 with $Y = 0$ out of 100? Set up the expression and say roughly how surprised you should be.

(d) Based on the "very high" group, are the model's probabilities *overconfident* or *underconfident*? Explain in a sentence.

(e) A well-calibrated model and an accurate model are not the same thing. Give an example of a classifier that is perfectly calibrated but has terrible accuracy.

Problem 6: When Accuracy Lies — Precision and Recall

You build a classifier to flag fraudulent transactions. On a test set of 200 transactions, the confusion matrix is:

	Predicted 1	Predicted 0
Actually 1	40	20
Actually 0	10	130

(a) Compute accuracy, precision $= \frac{TP}{TP+FP}$, and recall $= \frac{TP}{TP+FN}$.

(b) What accuracy would the "always predict 0" baseline achieve? What are its precision and recall?

(c) You lower the classification threshold from 0.5 to 0.3. Which way does recall move? Which way does precision (typically) move? Why?

(d) For fraud detection, which of precision and recall would you prioritize? Now answer the same question for a screening test that decides whether a patient gets an expensive and invasive follow-up procedure. Justify each answer in one sentence.

Problem 7: Is the Model Fair?

A loan-approval model is evaluated on two demographic groups. Let D be the protected demographic, G the model's guess, and T the true label.

Group	Applicants	Predicted $G=1$	Of those, truly $T=1$
$D=1$	200	60	48
$D=0$	200	100	70

(a) Compute $P(G = 1 \mid D = 1)$ and $P(G = 1 \mid D = 0)$. Does the model satisfy **parity**? Compute the ratio and check it against the US legal "80% rule" for disparate impact.

(b) Compute $P(T = 1 \mid G = 1, D = 1)$ and $P(T = 1 \mid G = 1, D = 0)$. Does the model satisfy **calibration** with respect to D, using the relaxed standard $\varepsilon = 0.2$?

(c) You have just shown that a model can pass one fairness criterion and fail the other. What does that tell you about the phrase "this model is fair"?

(d) "Fairness through unawareness" means simply deleting the protected attribute from the feature set. Explain, using the Facebook lookalike-audience result from lecture, why this often fails.

Problem 8: Decision Tree Entropy (pset7: decision_tree_entropy)

A DecisionTree has been trained on the heart-train dataset. Its root splits on the feature "C.4": datapoints with C.4 = 0 go left, C.4 = 1 go right. In the training dataset:

- There are 80 datapoints; 40 have Label 1 and the rest have Label 0.
- 52 points go left; of those, 17 have Label 1.
- 28 points go right; of those, 23 have Label 1.

Calculate the reduction in entropy achieved by this root node: (1) the entropy of the labels before the split, (2) the entropy of the labels in each child, (3) the *expected* entropy of the child a randomly selected training point lands in, and (4) the reduction, which is (1) minus (3). Model the labels in each node as i.i.d. Bernoulli and estimate p with the MLE.

Challenge Problem (optional): Platt Recalibration

(problem from the 2024 Winter Final)

You have an uncalibrated binary classifier that outputs $\hat{p} \in [0, 1]$. To fix it you use Platt Recalibration:

$$P(Y = 1 \mid \hat{p}) = \sigma(a \cdot \hat{p} - 0.5),$$

where a is the single parameter of the recalibration model. You are given the derivative

$$\frac{\partial}{\partial a}\sigma(a\hat{p} - 0.5) = \sigma(a\hat{p} - 0.5)\,[1 - \sigma(a\hat{p} - 0.5)]\,\hat{p}.$$

(a) For a new datapoint the uncalibrated model outputs $\hat{p} = 0.9$. With $a = 2$, what is the recalibrated probability that $Y = 1$?

(b) You are given n pairs $(\hat{p}^{(i)}, y^{(i)})$ where $y^{(i)} \in \{0, 1\}$ is the true outcome. Explain how you would estimate the value of a that makes the $y^{(i)}$ values as likely as possible, and solve for all partial derivatives your answer requires.

(c) Why must this recalibration model be fit on *held-out* data rather than the training set of the original classifier?