Lecture 1d: Decision Theory

From posterior probabilities to optimal actions

Jue Guo

University at Buffalo

May 7, 2026

Outline

Where we are

So far. Probability theory (1b) gave us a calculus for uncertainty. Cross-validation and the curse (1c) told us how to choose models in spite of finite data and high dimensions.

What's missing. A probability distribution over labels is not yet a decision. The cancer-screening machine has to commit to "treat" or "don't treat".

Decision theory – the topic of today – bridges probabilities and actions. Once we have $p(\mathbf{x}, t)$, making the optimal decision is, as Bishop puts it, "generally very simple, even trivial".

Two-stage view

Inference: from data, learn $p(\mathbf{x}, t)$ (or a piece of it).
Decision: from probabilities, choose an action.

A motivating example: cancer screening

- Input $\mathbf{x}$: pixel intensities of an X-ray image.
- Two classes: C_1 = "has cancer", C_2 = "no cancer".
- Joint distribution $p(\mathbf{x}, C_k)$ summarises everything.

Bayes' theorem (§1.2) gives the posterior

$$p(C_k \mid \mathbf{x}) \;=\; \frac{p(\mathbf{x} \mid C_k)\, p(C_k)}{p(\mathbf{x})}. \tag{1}$$

- $p(C_k)$: prior – prevalence of cancer in the population.
- $p(\mathbf{x} \mid C_k)$: likelihood – what an X-ray looks like under each class.
- $p(C_k \mid \mathbf{x})$: posterior – the revised belief given this patient's image.

Decision theory tells us what to do once we have these posteriors.

Decision regions and decision boundaries

A decision rule partitions the input space into decision regions $\mathcal{R}_k$ – one per class – such that any $\mathbf{x} \in \mathcal{R}_k$ is assigned to C_k.

Boundaries between regions are called decision boundaries (or *decision surfaces*).

Two-class probability of error.

$$p(\text{mistake}) = \int_{\mathcal{R}_1} p(\mathbf{x}, C_2)\, d\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, C_1)\, d\mathbf{x}. \tag{2}$$

Each $\mathbf{x}$ contributes to exactly one of the two integrals depending on which region we put it in. To minimize the sum, assign each $\mathbf{x}$ to the region whose integrand is *smaller* at that $\mathbf{x}$.

Bayes optimal classifier

From the product rule $p(\mathbf{x}, C_k) = p(C_k \mid \mathbf{x})\, p(\mathbf{x})$ and the fact that $p(\mathbf{x}) > 0$ is common, the rule simplifies:

$$\hat{k}(\mathbf{x}) \;=\; \arg\max_k \; p(C_k \mid \mathbf{x}). \tag{3}$$

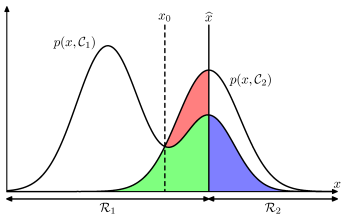

Figure 1.24

Optimal boundary x_0 sits where the joint densities cross. Shaded areas are the irreducible misclassification mass at this boundary; any other boundary only adds more. (PRML Fig. 1.24.)

K classes – maximize correctness

For K classes it is easier to maximize the probability of being *correct*:

$$p(\text{correct}) \;=\; \sum_{k=1}^{K} \int_{\mathcal{R}_k} p(\mathbf{x}, C_k)\, d\mathbf{x}. \tag{4}$$

Maximizing pointwise: assign each $\mathbf{x}$ to the class with the largest $p(\mathbf{x}, C_k)$, equivalently the largest posterior $p(C_k \mid \mathbf{x})$.

Bayes-optimal rule (0/1 loss)

$$\hat{k}(\mathbf{x}) \;=\; \arg\max_{k}\; p(C_k \mid \mathbf{x}).$$

Any other rule has a strictly larger error rate. This is the benchmark we compare every classifier against.

When mistakes aren't created equal

Cancer screening: missing a tumour kills a patient; a false alarm is a follow-up scan. The two errors should not have the same cost.

Loss matrix. L_{kj} is the cost when the true class is C_k and we decide C_j.

	decide cancer	decide normal
true cancer	0	1000
true normal	1	0

(PRML Fig. 1.25.) The asymmetry says: missing a true cancer is 1000 times worse than a false alarm. The optimal rule must shift the decision boundary toward declaring "cancer" more aggressively.

Minimum expected loss is still trivial

True class is unknown; average the loss under $p(\mathbf{x}, C_k)$:

$$\mathbb{E}[L] = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj}\, p(\mathbf{x}, C_k)\, d\mathbf{x}. \tag{5}$$

Pointwise minimisation, again using $p(\mathbf{x}, C_k) = p(C_k \mid \mathbf{x})\, p(\mathbf{x})$: assign $\mathbf{x}$ to the class j minimising

$$\sum_k L_{kj}\, p(C_k \mid \mathbf{x}). \tag{6}$$

Take-home

Once we know the posteriors $p(C_k \mid \mathbf{x})$, updating the loss matrix is free – we just multiply and pick the smallest. Decoupling inference from decision is what gives Bayesian classifiers their flexibility.

Concrete example – cancer threshold under L_{kj}

Use the loss matrix from before: $L_{12} = 1000$ (miss a cancer), $L_{21} = 1$ (false alarm), $L_{11} = L_{22} = 0$. Let $p \equiv p(C_1 \mid \mathbf{x})$ be the posterior probability of cancer.

Expected loss of each action.

$$\text{decide “cancer”}: \quad \sum_k L_{k1}\, p(C_k \mid \mathbf{x}) = 0 \cdot p + 1 \cdot (1-p) = 1 - p,$$

$$\text{decide “normal”}: \quad \sum_k L_{k2}\, p(C_k \mid \mathbf{x}) = 1000 \cdot p + 0 \cdot (1-p) = 1000\, p.$$

Decide “cancer” iff $1 - p \;<\; 1000\, p \iff p > \frac{1}{1001} \approx 0.001.$

	$p = 0.001$	$p = 0.05$	$p = 0.50$
L if “cancer”	0.999	0.95	0.50
L if “normal”	1.000	50.0	500
Optimal action	cancer (tie)	cancer	cancer

Asymmetric loss pushes the boundary from $p = 0.5$ (under 0/1 loss) to $p \approx 0.001$; the 1000:1 cost ratio is exactly the odds threshold.

Refusing to decide

Errors cluster where the largest posterior is barely larger than the next. Sometimes it's wiser to abstain.

Reject rule. Pick a threshold θ; reject (don't classify) whenever

$$\max_k p(C_k \mid \mathbf{x}) \;\leq\; \theta.$$

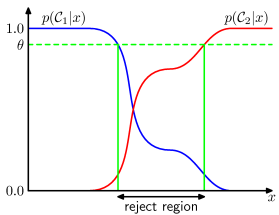

Figure 1.26

Shaded band: inputs whose largest posterior is below θ – abstain. $\theta = 1$ rejects everything; $\theta < 1/K$ rejects nothing. Easily extended to a loss matrix. (PRML Fig. 1.26.)

Three ways to build a classifier

Given training data, there are three increasingly direct strategies:

(a) Generative. Model $p(\mathbf{x} \mid C_k)$ and $p(C_k)$; combine via Bayes' theorem to get $p(C_k \mid \mathbf{x})$.
(b) Discriminative. Model the posterior $p(C_k \mid \mathbf{x})$ directly.
(c) Discriminant function. Learn $f(\mathbf{x})$ that maps inputs straight to labels. No probabilities.

Increasing simplicity, decreasing modelling assumptions.

Examples you'll meet

Naive Bayes, GMMs (a). Logistic regression, neural nets with softmax (b). Perceptron, SVM, k-NN with hard rules (c).

Trade-offs at a glance

	(a) Generative	(b) Discriminative	(c) Discriminant
Models	$p(\mathbf{x} \mid C_k),\ p(C_k)$	$p(C_k \mid \mathbf{x})$	$f : \mathbf{x} \to$ label
Data demands	High (fit $\mathbf{x}$-density)	Lower	Lowest
Outlier detection	Free, via $p(\mathbf{x})$	Limited	No
Reject / risk revision	Easy	Easy	Hard – retrain
Prior compensation	Easy (re-weight)	Easy if posterior	Not possible
Modular combination	Bayes + indep. assumption	Same	No probabilistic combination
Fits in high D	Hardest	Often easiest	Easiest; loses info

Posteriors are *useful even if you ultimately just want a label*: they let you change the loss, reject, re-weight class priors, or combine modules later without retraining.

Squared loss ⇒ the conditional mean

For regression with a real-valued target t, the expected squared loss is

$$\mathbb{E}[L] \;=\; \iint \{y(\mathbf{x}) - t\}^2 \, p(\mathbf{x}, t) \, d\mathbf{x} \, dt. \tag{7}$$

Take a functional derivative with respect to $y(\mathbf{x})$ and set to zero:

$$\frac{\delta \mathbb{E}[L]}{\delta y(\mathbf{x})} \;=\; 2 \int \{y(\mathbf{x}) - t\} \, p(\mathbf{x}, t) \, dt \;=\; 0. \tag{8}$$

Solving and using $p(\mathbf{x}, t) = p(t \mid \mathbf{x}) \, p(\mathbf{x})$,

$$\boxed{y^\star(\mathbf{x}) \;=\; \mathbb{E}[t \mid \mathbf{x}] \;=\; \int t \, p(t \mid \mathbf{x}) \, dt.} \tag{9}$$

The optimal regression function under squared loss is the conditional mean of t given $\mathbf{x}$.

Picture and decomposition

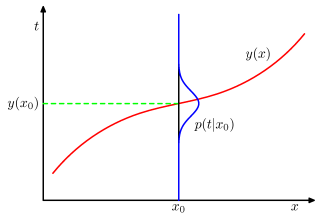

Figure 1.28

Decomposing $\{y(\mathbf{x}) - t\}^2$ around the conditional mean and taking expectations:

$$\mathbb{E}[L] \;=\; \int \{y(\mathbf{x}) - \mathbb{E}[t \mid \mathbf{x}]\}^2 \, p(\mathbf{x}) \, d\mathbf{x} \;+\; \int \mathrm{var}[t \mid \mathbf{x}] \, p(\mathbf{x}) \, d\mathbf{x}.$$

- First term: predictor's miss from the conditional mean – driven to zero with enough data and capacity.
- Second term: irreducible noise – the spread of t at each $\mathbf{x}$; no model removes it. (PRML Fig. 1.28; full bias-variance-noise decomposition in Ch. 3.)

Derivation – where the noise term comes from

Step 1. Add and subtract the conditional mean inside the squared error:

$$\{y(\mathbf{x}) - t\}^2 \;=\; \{[y(\mathbf{x}) - \mathbb{E}[t \mid \mathbf{x}]] + [\mathbb{E}[t \mid \mathbf{x}] - t]\}^2.$$

Step 2. Expand the square:

$$\{y(\mathbf{x}) - \mathbb{E}[t \mid \mathbf{x}]\}^2 \;+\; 2\,\{y(\mathbf{x}) - \mathbb{E}[t \mid \mathbf{x}]\}\{\mathbb{E}[t \mid \mathbf{x}] - t\} \;+\; \{\mathbb{E}[t \mid \mathbf{x}] - t\}^2.$$

Step 3. Take $\mathbb{E}_t[\cdot \mid \mathbf{x}]$, the inner expectation over t for fixed $\mathbf{x}$. The cross term vanishes, because

$$\mathbb{E}_t\Big[\{y(\mathbf{x}) - \mathbb{E}[t \mid \mathbf{x}]\}\{\mathbb{E}[t \mid \mathbf{x}] - t\} \,\Big|\, \mathbf{x}\Big] \;=\; \{y(\mathbf{x}) - \mathbb{E}[t \mid \mathbf{x}]\} \cdot \underbrace{\big(\mathbb{E}[t \mid \mathbf{x}] - \mathbb{E}_t[t \mid \mathbf{x}]\big)}_{=0} = 0.$$

Step 4. What remains, integrated over $\mathbf{x}$ under $p(\mathbf{x})$:

$$\mathbb{E}[L] \;=\; \underbrace{\int \{y(\mathbf{x}) - \mathbb{E}[t \mid \mathbf{x}]\}^2\, p(\mathbf{x})\, d\mathbf{x}}_{\text{model error}} \;+\; \underbrace{\int \operatorname{var}[t \mid \mathbf{x}]\, p(\mathbf{x})\, d\mathbf{x}}_{\text{noise floor}}.$$

The noise floor limits *any* estimator. Ch. 3 splits the model-error term into bias and variance.

Takeaways

- Decision = inference + a rule for turning probabilities into actions.
- Under 0/1 loss, the optimal classifier is $\arg\max_k p(C_k \mid \mathbf{x})$ – the Bayes optimal rule.
- Under a loss matrix L_{kj}, the rule generalises to $\arg\min_j \sum_k L_{kj}\, p(C_k \mid \mathbf{x})$.
- A reject option is just a threshold on the largest posterior – cheap to add, useful in safety-critical applications.
- Three classifier paradigms: generative, discriminative, discriminant. Posteriors are valuable even when you only need a label – they enable risk revision, rejection, prior compensation, and modular combination.
- For regression under squared loss, the optimal predictor is the conditional mean $\mathbb{E}[t \mid \mathbf{x}]$.

Reading

Bishop, PRML (2006), §1.5.