CMPUT 267: Machine Learning I · Fall 2026

Conditional Probability, Expectation, and Variance

Xi Ye

2026-09-10

Recall: Joint and Marginal Distributions

$X = 1$ means a review contains "great". $Y = 1$ means its sentiment is positive.

	$Y = 0$ Negative	$Y = 1$ Positive	$p_X(x)$
$X = 0$ No "great"	0.42	0.08	0.50
$X = 1$ Contains "great"	0.18	0.32	0.50
$p_Y(y)$	0.60	0.40	1.00

- The **joint PMF** $p_{X,Y}(x, y)$ assigns probability to a pair of values.
- A **marginal PMF** sums out the other variable. Using marginal probability, we know $P(Y = 1) = 0.4$ (review is positive) in general.

Recall: Joint and Marginal Distributions

$X = 1$ means a review contains "great". $Y = 1$ means its sentiment is positive.

	$Y = 0$ Negative	$Y = 1$ Positive	$p_X(x)$
$X = 0$ No "great"	0.42	0.08	0.50
$X = 1$ Contains "great"	0.18	0.32	0.50
$p_Y(y)$	0.60	0.40	1.00

We know $P(Y = 1) = 0.4$ (in general, the probabilitiy of a review being positive is 0.4)

But if we already observed that a review contains the word "great" ?? (say we know $X = 1$)

Intuitively, the sentiment is more likely to be positive.

This Lecture: Probability, Continued

- **Conditional probability**
- **Independence**
- **Expectation and variance**
- **Naive bayes** (Optional: For your interest)

This is a review class; Please read the course note

Conditional Distributions

A **conditional distribution** describes the possible values of one random variable after another variable has been observed.

For a joint distribution $p_{X,Y}$ over random variables X and Y, $P_{Y|X}(Y = y \mid X = x)$ denotes the probability of $Y = y$ given that we have observed $X = x$.

For the review sentiment classification example, $P(Y = 1 \mid X = 1)$ denotes the probabilitiy of the review being positive when we observed it contains “great”.

Example: Conditional Probability

	$Y = 0$ Negative	$Y = 1$ Positive	$p_X(x)$
$X = 0$ No “great”	0.42	0.08	0.50
$X = 1$ Contains “great”	0.18	0.32	0.50
$p_Y(y)$	0.60	0.40	1.00

Before observing X, the probability that a review is positive is the marginal probability

$$P(Y = 1) = 0.40.$$

Example: Conditional Probability

Suppose we observe $X = 1$.

	$Y = 0$ Negative	$Y = 1$ Positive	**Retained mass**
$X = 1$ Contains “great”	0.18	0.32	0.50

Considering all the cases with $X = 1, Y = 1$ with a weight of 0.32 out of 0.5.
Renormalize the retained row:

$$P(Y = 1 \mid X = 1) = \frac{0.32}{0.18 + 0.32} = 0.64.$$

Before and After Observing X

Before observing the feature

$$P(Y = 1) = 0.40$$

Marginal probability of a positive review

After observing “great”

$$P(Y = 1 \mid X = 1) = 0.64$$

Conditional probability of a positive review

The feature is useful because observing it changes the label distribution.

Calculating Conditional Probability

To calculate conditional probabilitiy:

- only retain outcomes consistent with $X = x$
- renormalize the retained probabilities so they sum to 1.

For $p_X(x) > 0$, the **conditional PMF** of Y given $X = x$ is

$$p_{Y|X}(y \mid x) = \frac{p_{X,Y}(x, y)}{p_X(x)}.$$

Here x is fixed, while y ranges over the possible values of Y.

The denominator $p_X(x) = \sum_y p_{X,Y}(x, y)$ is the probability mass retained when $X = x$.

Reverse the Condition

The same joint probability can answer a different question:

$$P(X = 1 \mid Y = 1) = \frac{P(X = 1, Y = 1)}{P(Y = 1)} = \frac{0.32}{0.40} = 0.80.$$

Positive sentiment given “great”

$P(Y = 1 \mid X = 1) = 0.64$

“Great” given positive sentiment

$P(X = 1 \mid Y = 1) = 0.80$

⚠ The order matters: in general, $P(Y \mid X) \neq P(X \mid Y)$.

Conditional Densities

When X and Y are continuous, conditioning produces a **conditional PDF**.

For a fixed x with $p_X(x) > 0$,

$$p_{Y|X}(y \mid x) = \frac{p_{X,Y}(x, y)}{p_X(x)}, \qquad p_X(x) = \int_{\mathcal{Y}} p_{X,Y}(x, y)\, dy.$$

The result is a density over y:

$$\int_{\mathcal{Y}} p_{Y|X}(y \mid x)\, dy = 1, \qquad P(a \le Y \le b \mid X = x) = \int_a^b p_{Y|X}(y \mid x)\, dy.$$

Here $p_X(x)$ is a marginal density–not $P(X = x)$. Even though $P(X = x) = 0$, the conditional density is well defined by the density ratio whenever $p_X(x) > 0$.

Example: A Conditional Density

Suppose (X, Y) has joint density $p_{X,Y}(x, y) = 2$ on the triangle $0 \leq y \leq x \leq 1$, and density 0 elsewhere. **What is** $p_{Y|X}(y \mid x)$?

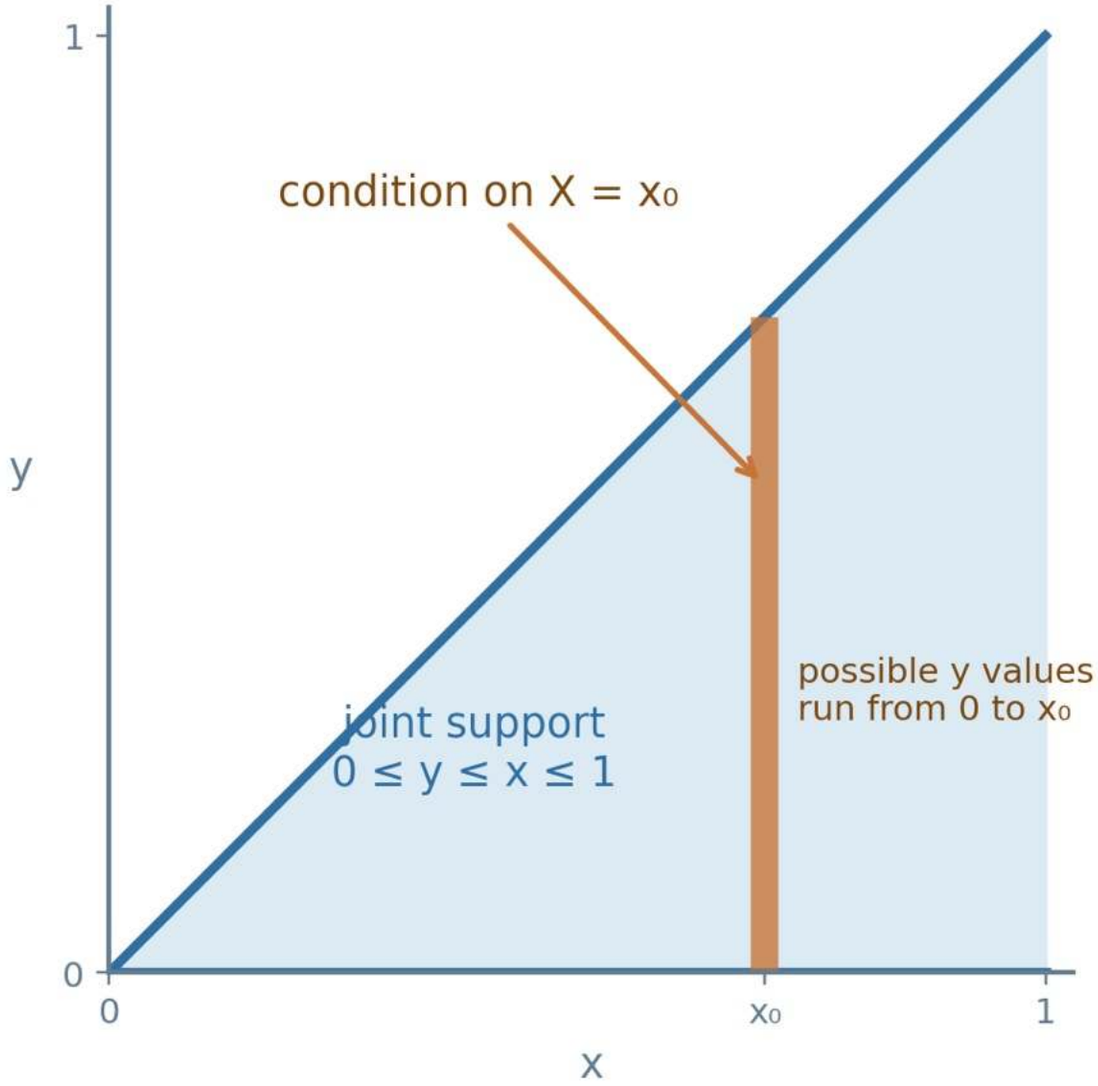
For $0 < x \leq 1$,

$$p_X(x) = \int_0^x 2\,dy = 2x.$$

Therefore,

$$p_{Y|X}(y \mid x) = \frac{2}{2x} = \frac{1}{x}, \qquad 0 \leq y \leq x.$$

Thus $Y \mid X = x \sim \text{Uniform}(0, x)$.



The Product Rule

Rearrange the definition of conditional probability:

$$p_{X,Y}(x, y) = p_{Y|X}(y \mid x)p_X(x) \qquad (1).$$

The same joint probability can also be written

$$p_{X,Y}(x, y) = p_{X|Y}(x \mid y)p_Y(y) \qquad (2).$$

Eq (1) and (2) describes the same joint probabilitiy.

Chain Rule for Several Variables

Using **Chain Rule** to describe the joint distribution of multiple (>2) random variable.

Apply the product rule repeatedly:

$$p(x_1, \ldots, x_d) = p(x_1) \prod_{j=2}^{d} p(x_j \mid x_1, \ldots, x_{j-1}).$$

For three variables,

$$p(x_1, x_2, x_3) = p(x_1) p(x_2 \mid x_1) p(x_3 \mid x_1, x_2).$$

Bayes' Rule Reverses a Conditional

Equate the two expressions for the joint probability:

$$p_{Y|X}(y \mid x)p_X(x) = p_{X|Y}(x \mid y)p_Y(y).$$

$$p_{Y|X}(y \mid x) = \frac{p_{X|Y}(x \mid y)p_Y(y)}{p_X(x)}$$

Bayes' Rule

$$p_{Y|X}(y \mid x) = \frac{p_{X|Y}(x \mid y)p_Y(y)}{p_X(x)}$$

Prior

$p_Y(y)$

What we believed before seeing x

Likelihood

$p_{X|Y}(x \mid y)$

How compatible x is with class y

Posterior

$p_{Y|X}(y \mid x)$

What we believe after seeing x

Bayes Rules allows us to invert conditional probabilities. Expressing $p_{X|Y}$ with $p_{Y|X}$ and the marginals (and vice versa).

Independence

Two random variables X and Y are **independent** if their joint distribution factorizes:

$$p_{X,Y}(x, y) = p_X(x)p_Y(y) \qquad \text{for every } x, y.$$

Whenever $p_X(x) > 0$, given $\frac{p_{X,Y}(x,y)}{p_X(x)} = p_{Y|X}(y \mid x)$, this is equivalent to

$$p_{Y|X}(y \mid x) = p_Y(y).$$

Independence means that observing X does not change the distribution of Y.

Examples: Independence and Dependence

Two fair coin flips

$$p_{X,Y}(x, y) = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2}$$

The first result does not update the distribution of the second: the flips are independent.

“Great” and review sentiment

$$0.32 \neq (0.50)(0.40)$$

Observing “great” changes the probability of a positive review: the feature and label are dependent.

Mixed Continuous and Discrete Variables

Suppose X is continuous and Y is discrete. Then $p_{X,Y}(x, y)$ is neither an ordinary PDF nor an ordinary PMF.

For $A \subseteq \mathcal{X}$ and $B \subseteq \mathcal{Y}$,

$$P(X \in A, Y \in B) = \int_A \sum_{y \in B} p_{Y|X}(y \mid x) p_X(x)\, dx.$$

For continuous variable, do integration; for discrete variable, use summation.

Example: Continuous Score, Discrete Label

Let $X \in [-1, 1]$ be a continuous sentiment score and $Y \in \{0, 1\}$ the review label.

Assume

$$p_X(x) = \frac{1}{2}, \qquad p_{Y|X}(1 \mid x) = \frac{x+1}{2}.$$

The probability of a nonnegative score and a positive label is

$$P(0 \le X \le 1, Y = 1) = \int_0^1 p_{Y|X}(1 \mid x) p_X(x)\, dx = \frac{3}{8}.$$

Expected Value

The **expected value**, or **mean**, is the long-run average obtained by repeatedly sampling a numerical random variable from its distribution.

Discrete

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x\, p(x)$$

Continuous

$$\mathbb{E}[X] = \int_{\mathcal{X}} x\, p(x)\, dx$$

The mean is not the most frequent value (which is called mode), and it need not be a value that X can actually take.

Example: Expected Value of a Die Roll

Let X be a fair six-sided die roll, so $p(x) = 1/6$ for $x \in \{1, \dots, 6\}$.

$$\begin{aligned} \mathbb{E}[X] &= \sum_{x=1}^{6} x\, p(x) \\ &= \frac{1+2+3+4+5+6}{6} = 3.5. \end{aligned}$$

> A die never shows 3.5. The expectation describes the average over many rolls, not one outcome.

Expectation of a Function

If $f : \mathcal{X} \to \mathbb{R}$, then $f(X)$ is numerical and has an expected value:

Discrete

$$\mathbb{E}[f(X)] = \sum_{x \in \mathcal{X}} f(x)p(x)$$

Continuous

$$\mathbb{E}[f(X)] = \int_{\mathcal{X}} f(x)p(x)\,dx$$

We can average $f(X)$ using the distribution of X; we do not first need to derive the distribution of $f(X)$.

Variance

The **variance** measures how far values typically spread from their mean.

$$\mathbf{Var}(X) = \mathbb{E}\big[(X - \mathbb{E}[X])^2\big].$$

A useful equivalent formula is

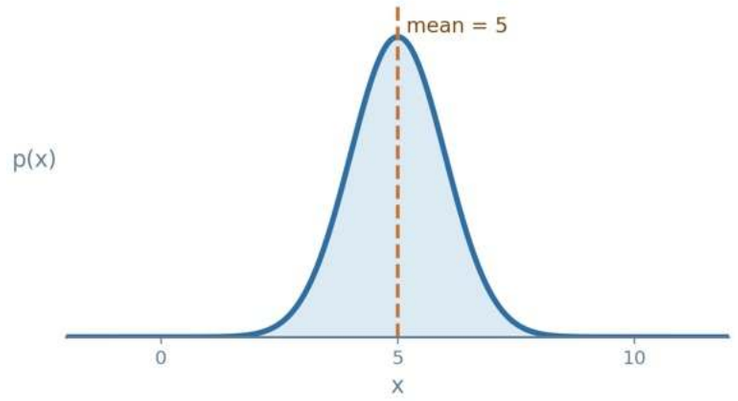
$$\mathbf{Var}(X) = \mathbb{E}[X^2] - \big(\mathbb{E}[X]\big)^2.$$

The standard deviation is $\sqrt{\mathbf{Var}(X)}$.

Same Mean, Different Variance

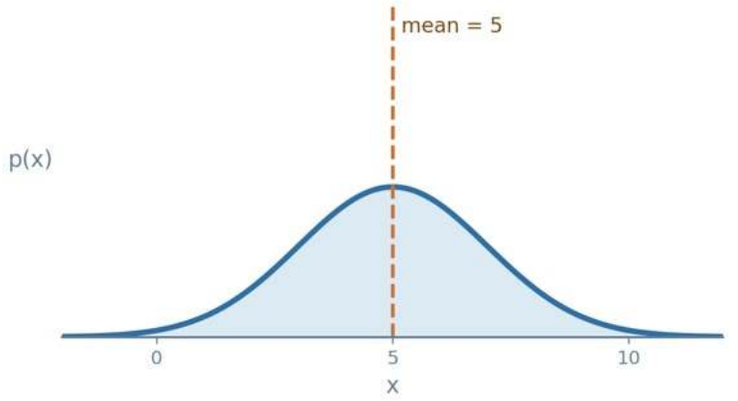
Both PDFs are centered at the same mean, $\mu = 5$, but their spreads are different.

Smaller Variance



$$X \sim \mathcal{N}(5, 1)$$
$$\mathrm{Var}(X) = 1$$

Larger Variance



$$X \sim \mathcal{N}(5, 4)$$
$$\mathrm{Var}(X) = 4$$

Useful Mean and Variance Results

Bernoulli

$$X \sim \text{Bernoulli}(\alpha)$$

$$\mathbb{E}[X] = \alpha$$
$$\text{Var}(X) = \alpha(1-\alpha)$$

Gaussian

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

$$\mathbb{E}[X] = \mu$$
$$\text{Var}(X) = \sigma^2$$

For a Gaussian, the two distribution parameters are exactly its mean and variance.

Some Properties of Expectation

For random variables X, Y and a constant c:

$$\begin{aligned}\mathbb{E}[cX] &= c\,\mathbb{E}[X],\\ \mathbb{E}[X+Y] &= \mathbb{E}[X] + \mathbb{E}[Y],\\ \mathbb{E}[\mathbb{E}[X \mid Y]] &= \mathbb{E}[X].\end{aligned}$$

If X and Y are independent, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.

Useful Properties of Variance

For a random variable X and a constant c:

$$\begin{aligned} \operatorname{Var}(c) &= 0, \\ \operatorname{Var}(cX) &= c^2 \operatorname{Var}(X), \\ X \perp Y &\implies \operatorname{Var}(X+Y) = \operatorname{Var}(X) + \operatorname{Var}(Y). \end{aligned}$$

Adding variances requires independence.

Recall: ML Pipeline

A Dataset Is a Random Variable

Before data collection, we do not know which examples will be sampled. Each example—and therefore the whole dataset—is random.

For one sampled example, let

$$Z_i = (\mathbf{X}_i, Y_i) \in \mathcal{X} \times \mathcal{Y}.$$

A dataset of n examples is the random tuple

$$D = (Z_1, \ldots, Z_n) \in (\mathcal{X} \times \mathcal{Y})^n.$$

Independent and Identically Distributed Data

A dataset is often modeled as **independent and identically distributed** (i.i.d.) examples $Z_1, \ldots, Z_n$.

Independent

Observing one sampled review does not change the distribution of another.

Identically Distributed

Every review is sampled from the same distribution $P_{\mathbf{X},Y}$.

$$(\mathbf{X}_i, Y_i) \overset{\text{iid}}{\sim} P_{\mathbf{X},Y}, \qquad i = 1, \ldots, n.$$

Predictors and Learners Are Random Variables

A learner is applied to a random dataset.

A predictor is applied to a random feature vector.

Application: Naïve Bayes

Classify with a probabilistic model

Naïve Bayes

Naïve Bayes is a probabilistic classifier that combines Bayes' rule with a conditional independence assumption.

For a review classifier:

- Y is the sentiment label: negative or positive;
- $\mathbf{X} = (X_1, \ldots, X_d)$ contains the observed features;
- the goal is to choose the most probable class after observing $\mathbf{X} = \mathbf{x}$.

Modeling Posterior

Each review has two binary features.

X_1

Review contains “great”

X_2

Review contains “boring”

After observing their values, the classifier wants

$$p_{Y|X_1,X_2}(y \mid x_1, x_2).$$

Estimating this posterior from counts requires enough labeled reviews for every exact feature pair (x_1, x_2).

Bayes Rewrites the Posterior

Bayes' rule lets us model it through a likelihood and a prior:

$$p_{Y|X_1,X_2}(y \mid x_1, x_2) = \frac{p_{X_1,X_2|Y}(x_1, x_2 \mid y)p_Y(y)}{p_{X_1,X_2}(x_1, x_2)}.$$

For fixed observed features (x_1, x_2), the denominator is the same for every candidate label. Therefore,

$$p_{Y|X_1,X_2}(y \mid x_1, x_2) \propto p_{X_1,X_2|Y}(x_1, x_2 \mid y)p_Y(y).$$

For classification, estimate the class prior and the feature likelihood. The normalizing denominator does not change which class wins.

Conditional Independence

X_1 and X_2 are conditionally independent given Y when

$$p(x_1, x_2 \mid y) = p(x_1 \mid y)p(x_2 \mid y).$$

Within positive reviews, knowing whether “great” appears does not change the probability of “boring”—and similarly within negative reviews.

Naïve Bayes Simplifies the Likelihood

Naïve Bayes treats the features as conditionally independent given the class:

$$p_{X_1,X_2|Y}(x_1, x_2 \mid y) \approx p_{X_1|Y}(x_1 \mid y) p_{X_2|Y}(x_2 \mid y).$$

For d features,

$$p_{\mathbf{X}|Y}(\mathbf{x} \mid y) \approx \prod_{j=1}^{d} p_{X_j|Y}(x_j \mid y).$$

We now estimate one simple feature distribution at a time within each class.

The Naïve Bayes Class Score

Combining Bayes' rule with the factorized likelihood gives

$$\hat{y} = \underset{y \in \mathcal{Y}}{\operatorname{argmax}} \; p_Y(y) \prod_{j=1}^{d} p_{X_j|Y}(x_j \mid y)$$

The model estimates the prior $p_Y(y)$ and one feature distribution $p_{X_j|Y}(x_j \mid y)$ at a time inside each class.

Example: A Review with Two Features

Observed review: $X_1 = 1$ (contains “great”) and $X_2 = 1$ (contains “boring”).

Quantity	**Negative:** $y = 0$	**Positive:** $y = 1$
$p_Y(y)$	0.60	0.40
$p_{X_1\mid Y}(1 \mid y)$	0.30	0.80
$p_{X_2\mid Y}(1 \mid y)$	0.70	0.20

Compare one score for each class.

Compare the Naïve Bayes Scores

For negative sentiment:

$$0.60 \times 0.30 \times 0.70 = 0.126.$$

For positive sentiment:

$$0.40 \times 0.80 \times 0.20 = 0.064.$$

Predict negative sentiment because $0.126 > 0.064$.

Optional: Naïve Bayes Summary

1. Bayes' rule expresses the posterior using a likelihood and a prior.
2. Naïve Bayes assumes features are conditionally independent given the class.
3. The assumption factorizes one joint likelihood into one-feature likelihoods.

Wrap-up

- Conditioning restricts and renormalizes a joint distribution.
- Bayes' rule reverses the direction of a conditional.
- Independence means an observation produces no update.
- Expectation describes the average value of a random quantity.
- Variance describes spread around the mean.
- Expected loss summarizes a predictor's performance on random examples.

This is a review class; Please read the course note