Introduction to Probabilistic Learning & Models

Handout Notes - Generative and Deep Learning (GDL)

Davide Bacciu - University of Pisa

Notation. Random variables are uppercase (e.g., X, Y, C) and observed values are lowercase (e.g., x, y, c). A dataset is $\mathcal{D} = \{x_1, \ldots, x_N\}$ with $N = |\mathcal{D}|$. We write $p(\cdot)$ for probability mass functions (discrete) and densities (continuous) whenever the meaning is clear from context. Conditional distributions are written $p(x \mid y)$; the log-likelihood is $\ell(\theta) = \log p(\mathcal{D} \mid \theta)$ when parameters θ are present.

1 Probabilistic learning models: what and why

Probabilistic learning models represent knowledge inferred from data *in the form of probability distributions*. This perspective is useful for at least three reasons:

- **Uncertainty is explicit.** Predictions are distributions, not just point estimates.
- **Generation is possible.** If a model specifies how data are distributed, we can sample synthetic data.
- **Prior knowledge can be incorporated.** By using priors on hypotheses or parameters, we can regularize learning and encode domain beliefs.

Modern ML tasks often involve many interacting variables. Writing an unrestricted joint distribution over all variables is typically infeasible: the parameter space grows exponentially with the number of variables, and inference (querying the model) can become computationally intractable. Graphical models address this by *factorizing* the joint distribution using conditional independences.

2 Graphical models: representation, inference, learning

A **graphical model** is a graph in which nodes represent random variables and edges represent probabilistic relationships. The graph provides a compact representation of a potentially huge joint distribution by encoding conditional independence assumptions.

2.1 Three recurring problems

Representation. Choose a structured family of distributions that can express relevant dependencies without exploding in complexity.

Inference. Given observations (evidence) $\mathbf{d}$, compute a query such as:

$$p(X \mid \mathbf{d}) \quad \text{or} \quad \arg\max_x p(x \mid \mathbf{d}).$$

This is the core "predict with the model" step.

Learning. Given data $\mathcal{D}$ and a fixed model structure, estimate parameters θ of p_θ. This can be framed as an inference problem over hypotheses/parameters.

2.2 Directed and undirected graphs (high-level view)

Different graph types correspond to different factorization patterns and conditional independence semantics:

- **Directed models (Bayesian networks):** edges suggest a directional factorization into conditional probabilities.
- **Undirected models (Markov random fields):** edges express soft constraints via potentials, factorizing the distribution into cliques.
- **Dynamic models:** structure evolves over time to represent stochastic processes.

3 Probability refresher

This section consolidates the probability tools that will be used throughout probabilistic learning.

3.1 Random variables and sample spaces

A **random variable** (RV) is a function from outcomes of a random experiment to a numerical (or categorical) value.

- **Discrete RV:** takes values in a finite or countably infinite set (e.g., $\{0, 1\}$, or $\{1, \dots, K\}$).
- **Continuous RV:** takes values in an uncountable set (e.g., $\mathbb{R}$).

We write X for the RV and x for a particular realization.

3.2 Probability mass functions and densities

For a discrete RV X with domain Ω,

$$p(X = x) \in [0, 1], \qquad \sum_{x \in \Omega} p(X = x) = 1.$$

For a continuous RV X with density $p(x)$,

$$p(x) \geq 0, \qquad \int_{\Omega} p(t)\, dt = 1, \qquad p(X \leq x) = \int_{-\infty}^{x} p(t)\, dt.$$

A common source of confusion: for continuous variables, $p(X = x) = 0$ for any single point x; probabilities are assigned to *intervals*.

3.3 Joint, marginal, and conditional distributions

Given two RVs X and Y, the joint distribution $p(x, y)$ specifies probabilities (or density mass) over pairs. Marginals are obtained by summing or integrating out other variables:

$$p(x) = \sum_{y} p(x, y) \quad \text{(discrete } Y\text{)}, \qquad p(x) = \int p(x, y)\, dy \quad \text{(continuous } Y\text{)}.$$

Conditionals are defined whenever $p(y) > 0$:

$$p(x \mid y) = \frac{p(x, y)}{p(y)}.$$

A conditional distribution is a *family* of distributions over x, indexed by the conditioning value y.

4 Sum rule, product rule, and the chain rule

Two rules power most manipulations in probabilistic models.

4.1 Product rule (definition of conditional probability)

$$p(x, y) = p(x \mid y)\, p(y) = p(y \mid x)\, p(x).$$

4.2 Sum rule (marginalization)

$$p(x) = \sum_{y} p(x, y) \quad \text{or} \quad p(x) = \int p(x, y)\, dy.$$

4.3 Chain rule

For n random variables $X_1, \ldots, X_n$ and any ordering,

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i \mid x_1, \ldots, x_{i-1}).$$

Worked example — Chain rule factorization for three variables

For X, Y, Z,

$$p(x, y, z) = p(x)\, p(y \mid x)\, p(z \mid x, y).$$

Equivalently (different ordering),

$$p(x, y, z) = p(z)\, p(x \mid z)\, p(y \mid x, z).$$

Both are exact identities; the modeling choice is how to *parameterize* the conditionals.

5 Bayes' rule and a "hypothesis" view of learning

Bayes' rule connects likelihood, prior, and posterior:

$$p(h \mid \mathbf{d}) = \frac{p(\mathbf{d} \mid h)\, p(h)}{p(\mathbf{d})}, \qquad p(\mathbf{d}) = \sum_{h' \in H} p(\mathbf{d} \mid h') p(h').$$

Here:

- $p(h)$ is a prior belief over hypotheses,
- $p(\mathbf{d} \mid h)$ is the likelihood of the observed data under hypothesis h,
- $p(h \mid \mathbf{d})$ is the posterior belief after observing data.

5.1 Three approaches to inference (and learning)

Suppose we want to predict a variable X after seeing data $\mathbf{d}$ and considering a hypothesis space H.

Bayesian prediction (posterior averaging).

$$p(X \mid \mathbf{d}) = \sum_{h \in H} p(X \mid h)\, p(h \mid \mathbf{d}).$$

MAP (maximum a posteriori). Choose the most probable hypothesis under the posterior:

$$h_{\text{MAP}} = \arg\max_{h \in H} p(h \mid \mathbf{d}) = \arg\max_{h \in H} p(\mathbf{d} \mid h)\, p(h).$$

ML (maximum likelihood). With uniform priors over H, MAP reduces to ML:

$$h_{\text{ML}} = \arg\max_{h \in H} p(\mathbf{d} \mid h).$$

6 Independence and conditional independence

Independence expresses the idea that learning one variable does not change beliefs about another.

6.1 Independence

X and Y are independent iff

$$p(x, y) = p(x)\, p(y),$$

equivalently $p(x \mid y) = p(x)$.

6.2 Conditional independence

X and Y are conditionally independent given Z iff

$$p(x, y \mid z) = p(x \mid z)\, p(y \mid z),$$

equivalently $p(x \mid y, z) = p(x \mid z)$.

Conditional independences are the central currency of graphical models: they are precisely what allows us to factorize large joint distributions into smaller pieces.

7 Expectation

The expectation of a function $f(X)$ under $p(x)$ is the average value of f weighted by the probability of each outcome.

7.1 Discrete and continuous

$$\mathbb{E}[f(X)] = \sum_x p(x)\, f(x) \quad \text{(discrete)}, \qquad \mathbb{E}[f(X)] = \int p(x)\, f(x)\, dx \quad \text{(continuous)}.$$

7.2 Key properties

Expectation is linear:

$$\mathbb{E}[af(X) + bg(X)] = a\,\mathbb{E}[f(X)] + b\,\mathbb{E}[g(X)].$$

If X and Y are independent, then

$$\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\,\mathbb{E}[g(Y)].$$

Worked example — Computing an expectation for a categorical variable

Let $X \in \{1, 2, 3\}$ with $p(X = 1) = 0.2$, $p(X = 2) = 0.5$, $p(X = 3) = 0.3$. Then

$$\mathbb{E}[X] = \sum_{x=1}^{3} x\, p(X = x) = 1 \cdot 0.2 + 2 \cdot 0.5 + 3 \cdot 0.3 = 2.1.$$

8 Common distributions and "matching" priors

Probabilistic modeling chooses distributions aligned with the data domain.

8.1 Examples of likelihoods (data models)

- **Bernoulli:** $X \in \{0, 1\}$ with $p(X = 1) = \lambda$.
- **Categorical:** $X \in \{1, \ldots, K\}$ with parameter vector $\boldsymbol{\lambda}$, $\sum_k \lambda_k = 1$.
- **Univariate Gaussian:** $X \in \mathbb{R}$ with parameters (μ, σ^2).
- **Multivariate Gaussian:** $\mathbf{X} \in \mathbb{R}^d$ with parameters $(\boldsymbol{\mu}, \Sigma)$.

8.2 Conjugate priors (why certain distributions appear together)

A prior is **conjugate** to a likelihood if the posterior has the same functional form as the prior. Conjugacy is valuable because it yields closed-form posteriors and interpretable updates.

Worked example — Beta–Bernoulli conjugacy (posterior update)

Let $X \in \{0, 1\}$ and $p(X = 1 \mid \lambda) = \lambda$ with parameter $\lambda \in (0, 1)$. Assume a Beta prior $p(\lambda) = \text{Beta}(\alpha, \beta) \propto \lambda^{\alpha-1}(1-\lambda)^{\beta-1}$. Given N i.i.d. observations with N_1 ones and N_0 zeros, the likelihood is

$$p(\mathcal{D} \mid \lambda) = \lambda^{N_1}(1-\lambda)^{N_0}.$$

Posterior is proportional to likelihood times prior:

$$p(\lambda \mid \mathcal{D}) \propto \lambda^{N_1}(1-\lambda)^{N_0}\, \lambda^{\alpha-1}(1-\lambda)^{\beta-1} = \lambda^{(\alpha+N_1)-1}(1-\lambda)^{(\beta+N_0)-1}.$$

Thus,

$$\boxed{p(\lambda \mid \mathcal{D}) = \text{Beta}(\alpha + N_1,\ \beta + N_0).}$$

A similar relationship holds between categorical/multinomial likelihoods and Dirichlet priors.

9 Inference and learning in probabilistic models

9.1 Inference: querying unknown variables

Inference asks: given observed variables (evidence), what can we say about unobserved variables? For example, in a student model, one might want $p(\text{graduate} \mid \text{exam}_1, \ldots, \text{exam}_n)$. In general, inference typically requires computing marginals or conditional marginals in a structured distribution.

9.2 Learning: estimating parameters as an inference problem

Learning can be viewed as a specific inference problem:

> Given observations $\mathbf{d}$ and a fixed model structure, infer parameters θ of p_θ that best explain the data.

In later lessons, "best" will be formalized via ML, MAP, or Bayesian learning.

10 Worked scenario: the candy box problem

To cement Bayes' rule as a learning tool, consider a finite hypothesis set $H = \{h_1, \ldots, h_5\}$ representing different candy-box compositions. Each hypothesis implies a probability of drawing

a cherry or lime candy. We observe a sequence of candies $\mathbf{d} = (d_1, \ldots, d_N)$ sampled i.i.d. (with replacement) from the unknown box.

10.1 Posterior computation

The posterior is

$$p(h_i \mid \mathbf{d}) \propto p(\mathbf{d} \mid h_i)\, p(h_i), \qquad p(\mathbf{d} \mid h_i) = \prod_{j=1}^{N} p(d_j \mid h_i).$$

This is a classic "prior times likelihood" update, repeated as evidence accumulates.

Worked example — Sequential Bayes updates for i.i.d. observations

Let $\mathbf{d}_{1:t} = (d_1, \ldots, d_t)$. Then

$$p(h \mid \mathbf{d}_{1:t}) \propto p(d_t \mid h)\, p(h \mid \mathbf{d}_{1:t-1}).$$

So the posterior after t observations is the previous posterior multiplied by the likelihood of the new observation. This is the basis of online belief updating.

10.2 Prediction

Once we have $p(h \mid \mathbf{d})$, Bayesian prediction averages over hypotheses:

$$p(d_{N+1} = \text{lime} \mid \mathbf{d}) = \sum_{i=1}^{5} p(d_{N+1} = \text{lime} \mid h_i)\, p(h_i \mid \mathbf{d}).$$

MAP prediction instead uses only the most probable hypothesis under $p(h \mid \mathbf{d})$.