MSE 546: Advanced Machine Learning

Sirisha Rambhatla

University of Waterloo

Lecture

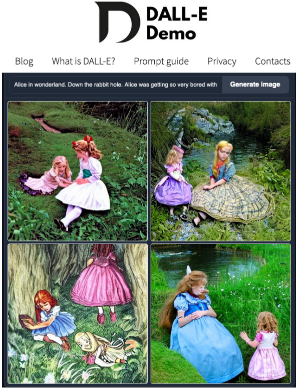
Math Background: Probability, Statistics and Information Theory

Outline

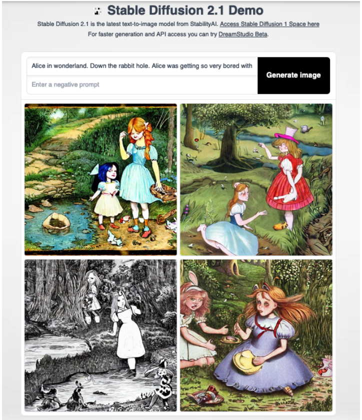
Outline

Conditional Generation of Data

Prompt: "Alice in wonderland. Down the rabbit hole. Alice was getting so very bored with her sister on the riverbank."



https://dalledemo.com/ [1, 2]



https://stablediffusion.fr/demo GitHub [3]

Probability: Terms and Notation

Sample Space: a set of all possible outcomes or realizations of some random trial.
Example: : Toss a coin twice; the sample space is $\Omega = \{HH, HT, TH, TT\}$.

Event: A subset of sample space
Example: : the event that at least one toss is a head is $A = \{HH, HT, TH\}$.

Probability: We assign a real number $P(A)$ to each event A, called the probability of A.
Example: In the coin tossing example

$$P(A) = \frac{3}{4}$$

Probability Axioms

The probability P must satisfy three axioms:

1. **Non-negativity:** $P(A) \geq 0$ for every A;
   - *Probabilities cannot be negative!*
2. **Unit Measure:**$P(\Omega) = 1$;
   - *Probability of entire samples space is 1*
   - i.e. "something happens" and that there are no events outside of the sample space
3. **Mutually exclusive events:** If $A_1, A_2, \ldots$ are disjoint, then $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$
   - **disjoint**: $A_1 \wedge A_2 = \emptyset$, $\wedge$ is logical AND.
   - $\cup$ is "union" which can be represented by $\vee$ logical OR
   - *Helps us to analyze events of interest and convert logical operations into arithmetic*

Probability: Terms and Notation

Probability of a union of two events: The probability of event A or B happening is given by:

$$P(A \vee B) = P(A) + P(B) - P(A \wedge B)$$

Random Variables

Definition: A random variable is a function that maps from a random event to a real number, i.e. $X : \Omega \to R$, that assigns a real number $X(\omega)$ to each outcome ω.

Example: In coin tossing, we let $H \to 1$ and let $T \to 0$. The event "at least one toss is a head" then can be written as $X_1 + X_2 > 0$, where X_1 and X_2 are the random variables (ie, 1, or 0 corresponding to the first toss and the second toss respectively).

Two Types: Discrete (e.g. Bernoulli in Coin toss) and Continuous (e.g. Gaussian)

From Random Variables to Data

Data The data are specific realizations of random variables

$$(X_1 = 1, X_2 = 0), (X_1 = 1, X_2 = 1), (X_1 = 0, X_2 = 0)$$

are 3 observations from the coin toss experiments (note that each experiment involves tossing twice).

Statistic: A statistic is any function of the data or random variables, e.g. mean, variance etc.

Distribution Function: Discrete Random Variable

Definition: Suppose X is a random variable and x is a specific value that it can take, then

For **discrete r.v.** X, the *probability mass function* is defined as

$$f_X(x) = P(X = x) : \text{probability of } X \text{ takes the value of } x$$

Example: For a fair coin $P(X = 1) = 1/2$, where X is either 0 ('T') or 1 ('H').

Distribution Function: Continuous Random Variable

Definition: Suppose X is a random variable and x is a specific value that it can take, then

For a **continuous r.v.** X, $f_X(x) \geq 0$ is the *probability density function*, for every $a \leq b$:

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

where $\int_{-\infty}^{\infty} f(x)dx = 1$. Note: for continuous distributions $P(X = x) = 0$

Cumulative distribution function (CDF) of X : $F_X(x) = P(X \leq x)$. If $F(x)$ is differentiable everywhere, $f(x) = F'(x)$.

Expectation

Expected Values

- Of a function $g(\cdot)$ of a **discrete r.v.** X,

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

Example: $E[g(X)]$ for tossing a fair coin is

$$\mu = 1 \times P(X = 1) + 0 \times P(X = 0) = 1/2$$

- Of a function $g(\cdot)$ of a **continuous r.v.** X,

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f(x) dx.$$

Expectation

Mean and Variance $\mu = E[X]$ is the mean; $var[X] = E[(X-\mu)^2]$ is the variance. Hence, we have $var[X] = E[X^2] - \mu^2$.

Example: The variance in our coin tossing example is

$$var[X] = E[(X-\mu)^2] = (1-1/2)^2 P(X=1) + (0-1/2)^2 P(X=0) = \frac{1}{4}$$

Linearity of Expectation: For r.v $X_1, \ldots, X_n$:

$$E[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} E[X_i].$$

Common Distributions

Discrete variable	Probability function	Mean
Uniform $X \sim U[1, \dots, N]$	$1/N$	$\frac{N+1}{2}$
Binomial $X \sim Bin(n, p)$	$\binom{n}{x} p^x (1-p)^{(n-x)}$	np
Geometric $X \sim Geom(p)$	$(1-p)^{x-1} p$	$1/p$
Poisson $X \sim Poisson(\lambda)$	$\frac{e^{-\lambda}\lambda^x}{x!}$	λ
Continuous variable	Probability density function	Mean
Uniform $X \sim U(a, b)$	1/ (b-a)	(a + b)/2
Gaussian $X \sim N(\mu, \sigma^2)$	$\frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2\sigma^2}(x-\mu)^2)$	μ
Gamma $X \sim \Gamma(\alpha, \beta)$ $(x \geq 0)$	$\frac{1}{\Gamma(\alpha)\beta^a} x^{a-1} e^{-x/\beta}$	$\alpha\beta$
Exponential $X \sim exponen(\beta)$	$\frac{1}{\beta} e^{-\frac{x}{\beta}}$	β

Multivariate Distributions

Dealing with two random variables

$f_{X,Y}(x, y) = P(X = x, Y = y)$: probability of X taking x *and* Y taking y

Example: Let X represent 'Vertical Jump Height' and Y represents players (either 'SN' or 'LJ').

$$P(X = 40 \text{ inches}, Y = `LJ')$$

probability that the vertical jump height is 40 inches AND the player is LJ

Multivariate Distributions: Marginal Distribution

Marginal distribution

$$P(X = x) = \sum_y P(X = x, Y = y), P(Y = y) = \sum_x P(X = x, Y = y)$$

Example: Represents the probability of vertical jump being x (irrespective of who the player is.)

Discrete Case:

$$f_X(x) = P(X = x) = \sum_y P(X = x, Y = y) = \sum_y f_{X,Y}(x, y)$$

Continuous Case

$$f_X(x) = \int_y f_{X,Y}(x, y)dy$$

Multivariate Distributions: Conditional Distribution

Conditional distribution

$$P(X = x|Y = y)$$

Represents the probability that vertical jump height is x GIVEN that the player is y.

$$P(Y = y|X = x)$$

Represents the probability that the player is y GIVEN that the vertical jump height is x.

Computed as

$$f_{X|Y}(x|y) = P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

Bayes Rule

Consider the **Conditional**:

$$P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}$$

This relationship can be used to derive the relationship between the two conditionals $P(X = x|Y = y)$ and $P(Y = y|X = x)$. This is **Bayes Rule**, i.e.

$$P(Y = y|X = x) = \frac{P(X = x, Y = y)}{P(X = x)} = \frac{P(X = x|Y = y)P(Y = y)}{P(X = x)}$$

Here we used the fact that
$P(X = x, Y = y) = P(X = x|Y = y)P(Y = y)$

Bayes Rule

Law of total Probability: Relates **conditional** to **marginal**

For the discrete case, say X takes values as $x_1, x_2, \dots$ the **marginal** can be written as

$$P(Y=y) = \sum_x P(X=x, Y=y) = \sum_x P(Y=y|X=x)P(X=x)$$

Therefore, we have

$$f_Y(y) = \sum_j f_{Y|X}(y|x_j) f_X(x_j)$$

Bayes Rule

(Simple Form)

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

(Discrete Random Variables)

$$f_{X|Y}(x_i|y) = \frac{f_{Y|X}(y|x_i)f_X(x_i)}{\sum_j f_{Y|X}(y|x_j)f_X(x_j)}$$

(Continuous Random Variables)

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)f_X(x)}{\int_x f_{Y|X}(y|x)f_X(x)dx}$$

Independence

Independent Variables X and Y are *independent* if and only if:

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

or $f_{X,Y}(x, y) = f_X(x)f_Y(y)$ for all values x and y.

IID variables: *Independent and identically distributed* (IID) random variables are drawn from the same distribution and are all mutually independent.

Correlation

Covariance

$$cov(X, Y) = E[(X - \mu_x)(Y - \mu_y)],$$

Correlation coefficients

$$corr(X, Y) = Cov(X, Y)/\sigma_x \sigma_y$$

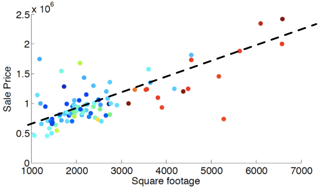
Independence $\Rightarrow$ Uncorrelated ($corr(X, Y) = 0$).

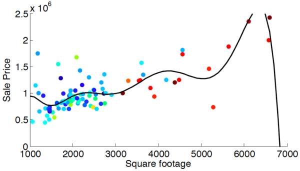
However, the reverse is generally not true.

Outline

Model fitting from Data

The process of estimating parameters $\boldsymbol{\theta}$ from $\mathcal{D}$ is called **model fitting**, or **training**, and is at the heart of machine learning.





Fitting a line (linear model) to Data. Fitting a polynomial (degree $d > 1$)

BUT

- *How do we know which model to fit?*
- *How do we learn the parameters of these models?*

Point Estimation

Definition The *point estimator* $\hat{\theta}_N$ is a function of samples $X_1, \ldots, X_N$ that approximates a parameter θ^* of the distribution of X_i. For example, Suppose $X_1, \ldots, X_n$ are random variables, you may have seen empirical estimates of mean as

Sample Mean

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i$$

Sample Bias: The bias of an estimator is

$$bias(\hat{\theta}_N) = E_\theta[\hat{\theta}_N] - \theta^*$$

An estimator is *unbiased estimator* if $E_\theta[\hat{\theta}_N] = \theta^*$

Why care about unbiasedness? If an estimator is *unbiased* then it gives you the *exact* parameter of interest (*in expectation*)!

Unbiased Estimators

Is unbiasedness enough?

Example As estimator that looks as only one datapoint is also unbiased.

$$\hat{\theta}(\mathcal{D}) = x_1$$

What is the problem with it? It will not *generalize* to new data!

Variance is also important

So the variance of an estimator is also important.

$$Var[\hat{\theta}] := E[\hat{\theta}^2] - \left(E[\hat{\theta}]\right)^2$$

How low can this go? This is answered by the celebrated **Cramer-Rao Lower Bound**. (Out of the scope of this course)

Bias-Variance Trade-off

A fundamental trade-off that needs to be made when picking a model, assuming that the goal is to minimize the mean squared error (MSE) of a estimate.

Given data samples $S = \{\boldsymbol{x}_n, y_n\}_{n=1}^N$ iid from the same distribution.

- y be the true label (observed, and deterministically related to $\boldsymbol{x}$)
- $f_S(\boldsymbol{x})$ denote the estimator or the model learned using S
- $\bar{f}(\boldsymbol{x}) = E[f_S(\boldsymbol{x})]$ be its expected value
- mean squared error (MSE) $:= E[(y - f_S(\boldsymbol{x}))^2]$

The for each fixed $\boldsymbol{x}$,

$$E[(y - f_S(\boldsymbol{x}))^2] = Var[f_S(\boldsymbol{x})] + (Bias(f_S(\boldsymbol{x})))^2 \textit{ Show?}$$

Take-away: Assuming our goal is to minimize squared error, it might be wise to use a biased estimator, so long as it reduces our variance by more than the square of the bias. *This holds for classical ML models and shallow neural networks (NNs) BUT has been found to not impact overparameterized NNs [4]!*

Maximum Likelihood Estimation

Motivation: Most machine learning training boils down to an optimization problem of the form

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \, \mathcal{L}(\boldsymbol{\theta})$$

Maximum Likelihood Estimation picks the parameters $\boldsymbol{\theta}$ that assign the highest probability to the training data.

$$\hat{\boldsymbol{\theta}}_{mle} := \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \, p(\mathcal{D}|\boldsymbol{\theta})$$

This is a **point estimate** since it is a single estimator.

Maximum Likelihood Estimation

We usually assume the training examples $\mathcal{D} = \{\boldsymbol{x}_n, \boldsymbol{y}_n\}$ for $n \in \{1, \dots, N\}$ are independently sampled from the same distribution (iid assumption), so the (conditional) likelihood $p(\mathcal{D}|\boldsymbol{\theta})$ becomes

$$p(\mathcal{D}|\boldsymbol{\theta}) := \prod_{n=1}^{N} p(\boldsymbol{y}_n|\boldsymbol{x}_n, \boldsymbol{\theta})$$

It is easier to work with **log likelihood**

$$\hat{\boldsymbol{\theta}}_{mle} := \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{n=1}^{N} \log p(\boldsymbol{y}_n|\boldsymbol{x}_n, \boldsymbol{\theta})$$

Maximum Likelihood Estimation

We prefer positing this as a *minimization* of the negative log likelihood ($NLL(\boldsymbol{\theta})$):

$$\hat{\boldsymbol{\theta}}_{mle} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\ NLL(\boldsymbol{\theta}) \tag{1}$$

$$= \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\ -\sum_{n=1}^{N} \log p(\boldsymbol{y}_n|\boldsymbol{x}_n, \boldsymbol{\theta}) \tag{2}$$

Aside: It can be shown that the MLE achieves the Cramer Rao lower bound, and hence has the smallest asymptotic variance of any unbiased estimator (asymptotically optimal).

Empirical Risk Minimization

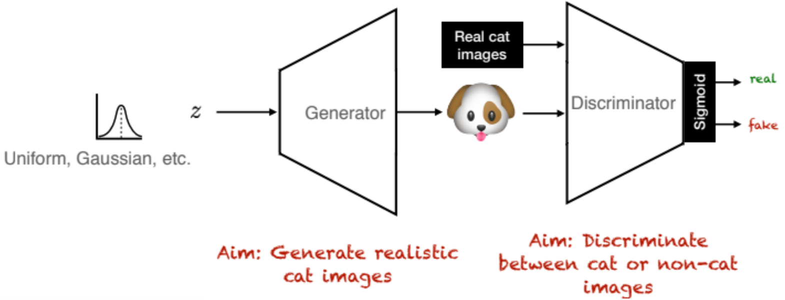
We can generalize MLE by replacing the (conditional) log loss term in $NLL(\boldsymbol{\theta})$ (1), with any loss function $\ell(\boldsymbol{y}_n, \boldsymbol{\theta}; \boldsymbol{x}_n)$ to get

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{n=1}^{N} \ell(\boldsymbol{y}_n, \boldsymbol{\theta}; \boldsymbol{x}_n) \tag{3}$$

This is known as **empirical risk minimization or ERM**, since it is the expected loss where the expectation is taken wrt the empirical distribution.

Outline

Learning to Generate Data



Learning in Generative Adversarial Networks

The Discrete Case: Shannon Entropy

Suppose X is a random variable which can have one of the m values: $x_1, \ldots, x_m.$, with probability $P(X = x_i) = p_i$ for $i \in m$.

Entropy (Shannon Entropy) is the average amount of *surprise* in a random variable's outcome.

$$H(X) = -\sum_{i=1}^{m} p_i \log_b p_i$$

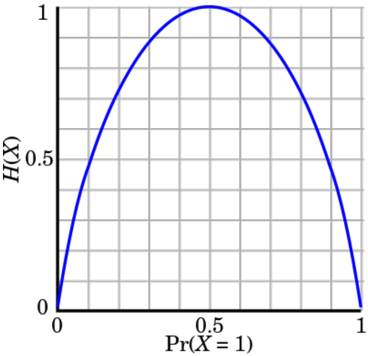
- "High entropy" means X is from a distribution closer to being uniform (more surprise);
- "Low entropy" means X is from varied (peaks and valleys) distribution (less surprise).
- Base "b" of the logarithm is usually 2, yielding units of $H(X)$ as bits.

Shannon Entropy: An Example

Entropy of a fair coin flip

- For a fair coin $p_i = 0.5$

$$
\begin{aligned}
H(X) &= -\textstyle\sum_{i=1}^{m} p_i \log_b p_i \\
&= -\sum_{i=1}^{2} 0.5 \log_2 0.5 \\
&= 1
\end{aligned}
$$



Entropy of coin flips

Entropy of a biased coin flip

- Say $p_1 = 0.8$

$$
\begin{aligned}
H(X) &= -\textstyle\sum_{i=1}^{m} p_i \log_b p_i \\
&= -0.8 \log_2 0.8 - 0.2 \log_2 0.2 \\
&= 0.7219
\end{aligned}
$$

Conditional Entropy

Conditional Entropy is the remaining entropy of a random variable Y given that the value of another random variable X is known.

$$\begin{aligned} H(Y|X) &= \sum_{i=1}^{m} p(X = x_i) H(p(Y|X = x_i)) \\ &= -\sum_{i=1}^{m} \sum_{j=1}^{n} p(x_i, y_j) \log p(y_j|x_i) \end{aligned}$$

This and related quantities are directly used in learning *Decision Trees*.

Cross Entropy

Cross Entropy the expected number of bits we need to represent a dataset coming from distribution p if we using distribution q.

$$H_{ce}(p,q) = -\sum_{i=1}^{m} p_i \log_b q_i$$

Kullback-Leibler (KL) Divergence

Kullback-Leibler divergence is a measure of distance between two distributions: a "true" distribution $p(X)$, and an arbitrary distribution $q(X)$.

$$\mathsf{KL}(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

Is there a relationship between KL Divergence and Cross-Entropy?

Outline

Reading

PiML1

- Chapter 2 – 2.1, 2.2 (2.2.1 - 2.2.4, 2.2.5.1 - 2.2.5.3), and 2.3.
- Chapter 3 – Eq (3.1), eq.(3.7), sections 3.1.3, and 3.1.4
- Chapter 4 – Section 4.1, 4.2.1, 4.2.2, 4.3 (intro), 4.5 (intro about MAP), 4.6 (intro), 4.7.6 (intro), 4.7.6.1, 4.7.6.2,
- Chapter 5 – Section 5.1.6.1
- Chapter 6 – 6.1 (6.1.1-6.1.4, 6.1.6. (intro)), 6.2 (6.2.1-6.2.2), 6.3 (6.3.1 - 6.3.2)
- https://ocw.mit.edu/courses/15-097-prediction-machine-learning-and-statistics-spring-dec694eb34799f6bea2e91b1c06551a0_MIT15_097S12_lec04.pdf

References I

[1] Alec Radford et al. "Learning Transferable Visual Models From Natural Language Supervision". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 18–24 Jul 2021, pp. 8748–8763. URL: https://proceedings.mlr.press/v139/radford21a.html.

[2] Aditya Ramesh et al. "Hierarchical text-conditional image generation with clip latents". In: *arXiv preprint arXiv:2204.06125* (2022).

[3] Robin Rombach et al. *High-Resolution Image Synthesis with Latent Diffusion Models*. 2021. arXiv: 2112.10752 [cs.CV].

References II

[4] Zitong Yang et al. "Rethinking Bias-Variance Trade-off for Generalization of Neural Networks". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, 13–18 Jul 2020, pp. 10767–10777. URL: https://proceedings.mlr.press/v119/yang20j.html.