Week1 Recitation

Zhenqiao/Zoey Song
CS190I Deep Learning

UC SANTA BARBARA

UC SANTA BARBARA
Computer Science

Outline

- **Questions?**
  - Deep learning? -- model architecture (neural network) & learning method
- **Probability Concepts**
- **Maximum Likelihood Estimation**
- **Maximum a Posterior Estimation**

Random Variable

- **Random Variable**
  - A mathematical formalization of an object which depends on random event. It's a mapping from possible outcomes in a sample space to a measurable space.
  - E.g.
    - Event: Flipping a coin
    - Sample space: the set {Head, Tail}
    - Possible outcomes: Head/Tail
    - Measurable space: {1, -1}

- **Probability Distribution**
  - Record the probabilities of all outcomes of a random variable X
  - E.g.: $P(X=1) = 0.5$, $P(X=-1)=0.5$

Discrete Random Variable

- **Sample Space**
  - A set of discrete values

- **Probability Mass Function**
  - The probability distribution of a discrete random variable is given by its probability mass function

  - A probability mass function should satisfy:

    - $\forall x \in X, 0 \leq P(x) \leq 1.$

    - $\Sigma_{x \in X} P(x) = 1$

Continuous Random Variable

- **Sample Space**
  - The random variable is valued in an interval of real numbers
- **Probability Density Function**
  - A function whose value at any given sample in the sample space describes a relative likelihood that the value of the random variable would be
  - PDF should satisfy:
    - The domain of PDF must be the set of all possible states of x
    - $\forall x \in X, p(x) \geq 0.$ Note we don't require p(x) <= 1
    - $\int p(x)dx = 1$
    -

Joint and Marginal Probability Distribution

- **Joint Distribution**
  - A probability distribution over multiple random variables
  - E.g., P(X=x, Y=y) denotes the probability that event X=x and Y=y happen simultaneously

- **Marginal Distribution**
  - For discrete variables, given the joint distribution P(X, Y), we can get the marginal distribution P(X) by the sum rule

$$P(\mathbf{x} = x) = \Sigma_y P(\mathbf{x} = x, \mathbf{y} = y)$$

  - For continuous random variables,

$$P(X=x) = \int_y P(X = x, Y = y)dy$$

Conditional Probability

- **Definition**
  - Probability of an event happens given another event

$$P(\mathbf{y} = y \mid \mathbf{x} = x) = \frac{P(\mathbf{y} = y, \mathbf{x} = x)}{P(\mathbf{x} = x)}$$

- Chain Rule (General Product Rule)

$$P(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(\mathbf{n})}) = P(\mathbf{x}^{(1)}) \prod_{i=2}^{n} P(x^{(i)} \mid x^{(1)}, \ldots, x^{(i-1)})$$

  - E.g., Language modeling

- **Independence**
  - If two random variables are independent, then their joint distribution equals to the product of their marginal distribution

$$P(\mathbf{x} = x, \mathbf{y} = y) = P(\mathbf{x} = x) P(\mathbf{y} = y)$$

Expectation

- **Definition**
  - The expectation of some function f(x) with respect to a probability distribution P(X) is the average value of f(x) when we take samples from P

- **For discrete random variables,**

$$\mathbb{E}_{\mathrm{x}\sim P}[f(x)] = \sum_x P(x)f(x)$$

- **For continuous random variables,**

$$\mathbb{E}_{\mathrm{x}\sim p}[f(x)] = \int p(x)f(x)dx$$

- **Expectations are linear,**

$$\mathbb{E}_{\mathrm{x}}[\alpha f(x) + \beta g(x)] = \alpha\mathbb{E}_{\mathrm{x}}[f(x)] + \beta\mathbb{E}_{\mathrm{x}}[g(x)]$$

Variance

- **Definition**
  - The variance gives a measure of how much the values of a function vary when we take samples from a probability distribution

$$\mathrm{Var}(f(x)) = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2\right]$$

  - The square root of the variance is called the standard deviation

Common Probability Distributions

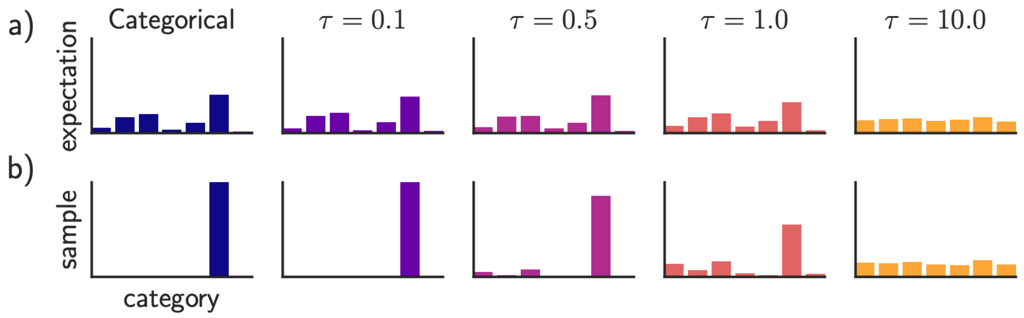
- **For discrete random variables**

  - **Bernoulli distribution**
    - A distribution over a single binary random variable, which is controlled by a single parameter **p**:

      P(X=1) = p, P(X=0) = 1 - p

    e.g. binary classification

  - **Categorical distribution**
    - Extends the above binary case to k states
    - E.g. multi-class classification

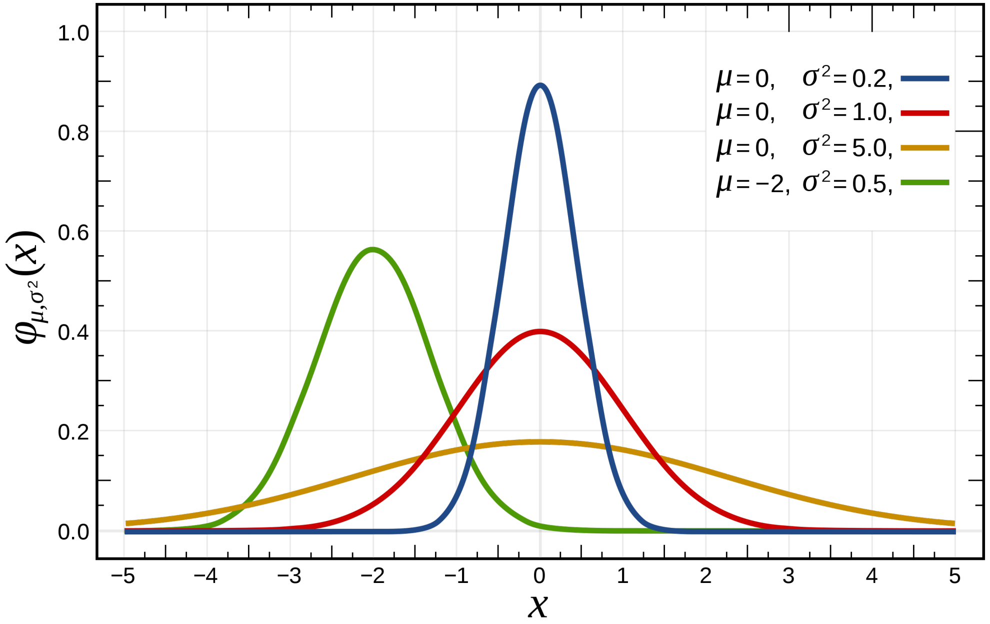


Common Probability Distributions

- **For continuous random variables**
  - **Gaussian distribution (Normal Distribution)**

$$\mathcal{N}(x;\mu,\sigma^2)=\sqrt{\frac{1}{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

  - **Expectation is μ**
  - **Variance is σ^2**



Bayes Rule

- **Definition**
  - Describes the probability of an event based on prior knowledge of conditions that might be related to that event

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

$$P(x \mid y) = \frac{P(x)P(y \mid x))}{P(y)}$$

$$P(x \mid y) = \frac{P(x)P(y \mid x)}{\Sigma_{x' \in X} P(x')P(y \mid x')}$$

- **E.g.**
  - Bayesian inference

Maximum Likelihood Estimation (MLE)

- **Definition**
  - $X_1, X_2, \ldots, X_N$ -- i.i.d random variables with probability distribution P(X| θ), where θ is the parammeter
  - Likelihood function L(x| θ) with a set of observations = {x, x_2, …, x_N}

$$L(X|\theta) = \prod_{i=1}^{N} P(x_i|\theta)$$

  - Then we can use MLE to find the empirically best θ **that maximizes L(x| θ)**

$$\hat{\theta} = argmax\, \mathbf{L(x|\ \theta)}$$

  - **For convenient computation,**

$$\hat{\theta} = argmax \log L(x|\theta) = argmax \sum_{i=1}^{N} \log P(x_i|\theta)$$

MLE Example

- **One-dimensional Gaussian distribution**

$$\mathcal{N}\left(x|\mu,\sigma^2\right) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

$$\ln p\left(\mathbf{x}|\mu,\sigma^2\right) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi).$$

- Setting the partial derivatives to 0, we can get

$$\mu_{ML} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\sigma_{ML}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_{ML})^2$$

- We can directly calculate the analytical solution for Gaussian distribution
- However, for more complicated functions such as neural networks (MLP, CNN, Transformer), there is no analytical solution. Usually, we can use gradient ascent to get the MLE solution
  - E.g., A language model such as GPT-2

MLE Example

- **One-dimensional Gaussian distribution: Note**

$$\mathrm{E}(\mu_{ML}) = \boldsymbol{E}\left(\frac{1}{N}\sum_{i=1}^{N} x_i\right) = \boldsymbol{\mu}$$

$$\boldsymbol{E}(\sigma_{ML}^2) = \mathbf{E}\left(\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_{ML})^2\right) = \frac{\boldsymbol{N-1}}{\boldsymbol{N}}\boldsymbol{\sigma^2}$$

- MLE systematically underestimates the variance of the distribution. This phenomenon is called bias
  - When N is large enough and in the limit N -> ∞, the bias is less significant
  - But when there are not enough samples (small N), the bias may be a serious problem
  - The issue of bias in maximum likelihood lies at the root of the over-fitting problem

MLE Example

- **Solve this problem**
  - Adding regularization to the parameter
    - E.g., L1 (Lasso) or L2 (Ridge) regularization
    - Dropout
  - Using Maximum Posterior estimation (MAP)

MAP

- **Description**
  - MAP can be used to obtain a point estimated of an unobserved quantity on the basis of empirical data. Different from MLE, it employs an augmented optimization objective which incorporates a prior distribution

$$\hat{\theta}_{MLE} = argmax\ P(\boldsymbol{x}|\theta)$$

$$\begin{aligned} \hat{\theta}_{MAP} &= \underset{\theta}{\operatorname{argmax}}\, P(\theta|\mathbf{x}) \\ &= \underset{\theta}{\operatorname{argmax}}\, \frac{P(\mathbf{x}|\theta)P(\theta)}{\int_\theta P(\mathbf{x}|\theta)P(\theta)d\theta} \\ &= \underset{\theta}{\operatorname{argmax}}\, P(\mathbf{x}|\theta)P(\theta) \end{aligned}$$

  - For Gaussian example, if we also use a Gaussian distribution for prior, then

$$\hat{\theta}_{MAP} = argmax\ \log f(x|\theta) - \frac{\theta^2}{2}$$

    - Equal to adding L2 regularization

  - Compared with Bayesian inference?

Any Question?