Probability Basics for Machine Learning

CSC411

*Based on many other people's slides and resource from Wikipedia.

Outline

- Motivation
- Notation, definitions, laws
- Probability distributions

Why Represent Uncertainty?

- The world is full of uncertainty
  - “What will the weather be like today?”
  - “Will I like this movie?”
  - “Is there a person in this image?”
- We’re trying to build systems that understand and (possibly) interact with the real world
- We often can’t *prove* something is true, but we can still ask how likely different outcomes are or ask for the most likely explanation

Why Use Probability to Represent Uncertainty?

- Write down simple, reasonable criteria that you'd want from a system of uncertainty (common sense stuff), and you always get probability.
- We will restrict ourselves to a relatively informal discussion of probability theory.

Notation

- A **random variable X** represents outcomes or states of the world.
- We will write p(x) to mean Probability(X = x)
- **Sample space**: the space of all possible outcomes (may be discrete, continuous, or mixed)
- p(x) is the **probability mass (density) function**
  - Assigns a number to each point in sample space
  - Non-negative, sums (integrates) to 1
  - Intuitively: how often does x occur, how much do we believe in x.

Joint Probability Distribution

- Prob(X=x, Y=y)
  - "Probability of X=x and Y=y"
  - $p(x, y)$

Conditional Probability Distribution

- Prob(X=x|Y=y)
  - "Probability of X=x given Y=y"
  - $p(x|y) = p(x,y)/p(y)$

Example

- Consider a bag of 3 red and 5 blue marbles
- We will take two marbles out of the bag, one at a time, without replacement.
  - Let X=(x1, x2) denote the event that the first marble has color x1 and the second marble has color x2. Both x1 and x2 can be R(Red) or B(Blue)
- Sample space: {(R,R), (R,B), (B,R), (B,B)}

Example

- Consider a bag of 3 red and 5 blue marbles
- We will take two marbles out of the bag, one at a time, without replacement.
  - Let X=R1 denote the event of drawing a red marble first, B1 a blue marble first, and R2, B2 of drawing a red and blue marble second respectively.
- Sample space: {R1,R2}, {B1,B2}, {R1,B2}, {B1,R2}

Example: Conditional Probability

P(x1,x2)	x1=R	x1=B
x2=R	3/28	15/56
x2=B	15/56	5/14

- Now assume x1=R
- What are the conditional probabilities P(x2=R|x1=R), P(x2=B|x1=R)?
  - P(x2=R|x1=R) = (3/28)/((3/28)+(15/56)) = (2/7)
  - P(x2=B|x1=R) = (15/56)/((3/28)+(15/56)) = (5/7)
- We “slice” the table by choosing column x1=R, and then renormalize
- Alternatively: What are P(R) and P(B) after we remove a red marble from the bag?

The Rules of Probability

- Sum Rule (marginalization/summing out):

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

$$p(x_1) = \sum_{x_2} \sum_{x_3} \cdots \sum_{x_N} p(x_1, x_2, \ldots, x_N)$$

- Product/Chain Rule:

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

$$p(x_1, \ldots, x_N) = p(x_1) p(x_2 \mid x_1) \ldots p(x_N \mid x_1, \ldots, x_{N-1})$$

Example: Marginalizing and Chaining

P(x1,x2)	x1=R	x1=B
x2=R	3/28	15/56
x2=B	15/56	5/14

- Marginalization:

  $P(x1=R) = P(R,R) + P(R,B) = (3/28) + (15/56) = 3/8$

- Chaining:

  $P(x2=R)$

  $= P(R,R)+P(B,R)$

  $= P(x2=R|x1=R)P(x1=R) + P(x2=R|x1=B)P(x1=B)$

  $= (2/7)*(3/8)+(3/7)*(5/8)=3/8$

Bayes' Rule

- One of the most important formulas in probability theory

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

- This gives us a way of "reversing" conditional probabilities

Independence

- Two random variables are said to be **independent** iff their joint distribution factors

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

$$p(x) = p(x\,|\,y) \quad \text{or} \quad p(y) = p(y\,|\,x)$$

- Two random variables are **conditionally independent** given a third if they are independent after conditioning on the third

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

Example: Independence

P(x1,x2)	x1=R	x1=B
x2=R	9/64	15/64
x2=B	15/64	25/64

- Now we sample **with replacement**
- The joint distribution has been changed accordingly
- P(R,B) = (3/8)*(5/8) = 15/64
- Notice that P(R,B) = P(x1=R)P(x2=B), the two trials are independent

Continuous Random Variables

- Outcomes are real values. Probability density functions define distribution.
  - E.g.,

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

- Probability mass in [a,b]

$$p(a \le x \le b) = \int_b^a \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}\mathrm{d}x$$

Continuous Random Variables

- Continuous joint distributions: replace sums with integrals, and everything holds
  - E.g., Marginalization and conditional probability

$$P(x,z) = \int_y P(x,y,z) = \int_y P(x,z \mid y)P(y)$$

Summarizing Probability Distributions

- It is often useful to give summaries of distributions without defining the whole distribution (E.g., mean and variance)
- Mean: $$E[x] = \langle x \rangle = \int_x x \cdot p(x)\, dx$$
- Variance: $$\mathrm{var}(x) = \int_x (x - E[x])^2 \cdot p(x)\, dx$$

$$= E[x^2] - E[x]^2$$

Example 1: Bernoulli

- Binary random variable $X \in \{0,1\}$
- p(heads) = μ $\mu \in [0,1]$
- Coin toss

$$p(x \mid \mu) = \mu^{x}(1-\mu)^{1-x}$$

- Mean: μ
- Variance: μ(1- μ)

Example 2: Multinomial

$X \in \{1,2,...,K\}$

$\mu_k \in [0,1], \sum_{k=1}^{K} \mu_k = 1$

- p(X=k|μ) = μ_k
- For a single observation – die toss
  - Sometimes called Categorical

$X = (x_1, x_2, ..., x_K) \quad x_i \in \{0, 1\} \quad p(x_1, x_2, ..., x_K|\mu) = \prod_{k=1}^{K} \mu_k^{x_k}$

- Marginal distribution: $p(x_k \mid \mu) = \mu_k^{x_k}(1-\mu_k)^{1-x_k}$
- Mean of x_k: μ_k
- Variance of x_k: $\mu_k(1-\mu_k)$
- This is a generalization of the Bernoulli distribution to a 1 of K distribution
- Note that the x_k's are not independent since must sum to 1 over all k's

Example 2: Multinomial

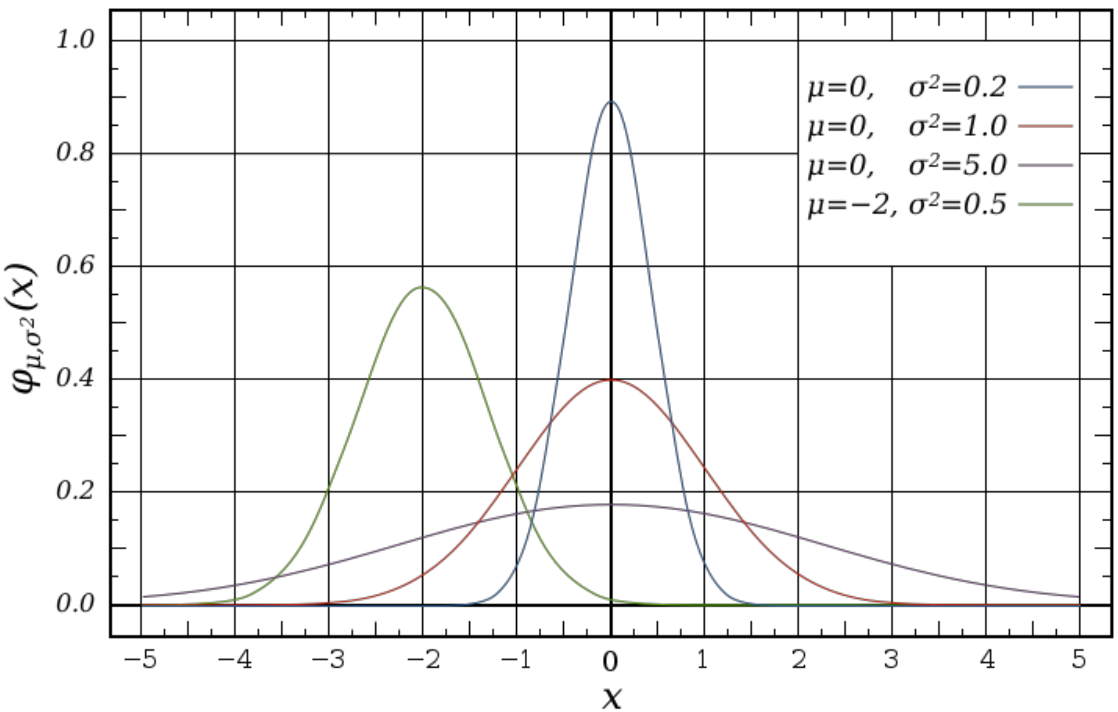
- For multiple observations $\mu_k \in [0,1], \sum_{k=1}^{K} \mu_k = 1$
  - integer counts on N trials
  - Prob(1 came out 3 times, 2 came out once,…,6 came out 7 times if I tossed a die 20 times)

$$\sum_{k=1}^{K} x_k = N \qquad P(x_1,\ldots,x_K \mid \mu) = \frac{N!}{\prod_k x_k!} \prod_{k=1}^{K} \mu_k^{x_k}$$

Example 3: Normal (Gaussian) Distribution

- Gaussian (Normal)

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



Example 3: Normal (Gaussian) Distribution

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

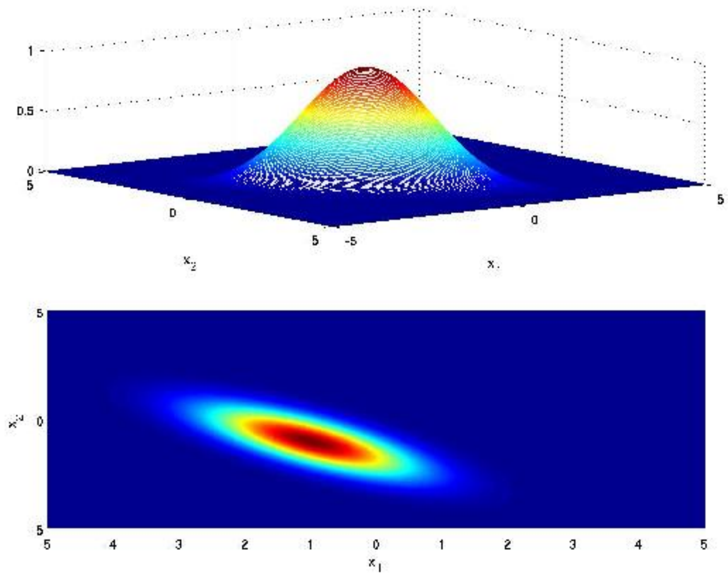
- μ is the mean
- σ^2 is the variance
- Can verify these by computing integrals. E.g.,

$$\int_{-\infty}^{\infty} x \cdot \frac{1}{\sqrt{2\pi\sigma}} \exp\left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\} dx = \mu$$

Example 3: Normal (Gaussian) Distribution

- Multivariate Gaussian

$$P(x \mid \mu, \Sigma) = \left|2\pi\Sigma\right|^{-1/2} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\}$$



Example 3: Normal (Gaussian) Distribution

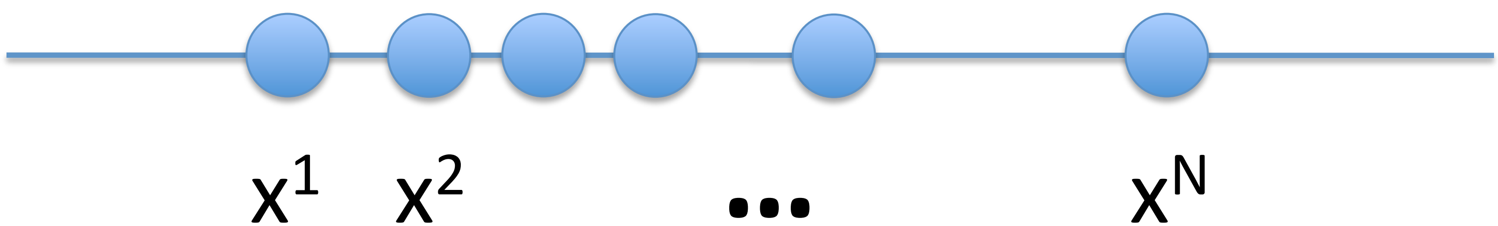
- Multivariate Gaussian

$$p(x \mid \mu, \Sigma) = \left|2\pi \Sigma\right|^{-1/2} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\}$$

- x is now a vector
- μ is the **mean vector**
- Σ is the **covariance matrix**

Example: Maximum Likelihood For a 1D Gaussian

- Suppose we are given a data set of samples of a Gaussian random variable X, D={x^1,…, x^N} and told that the variance of the data is σ^2



What is our best guess of μ?

*Need to assume data is independent and identically distributed (i.i.d.)

Example: Maximum Likelihood For a 1D Gaussian

What is our best guess of μ?

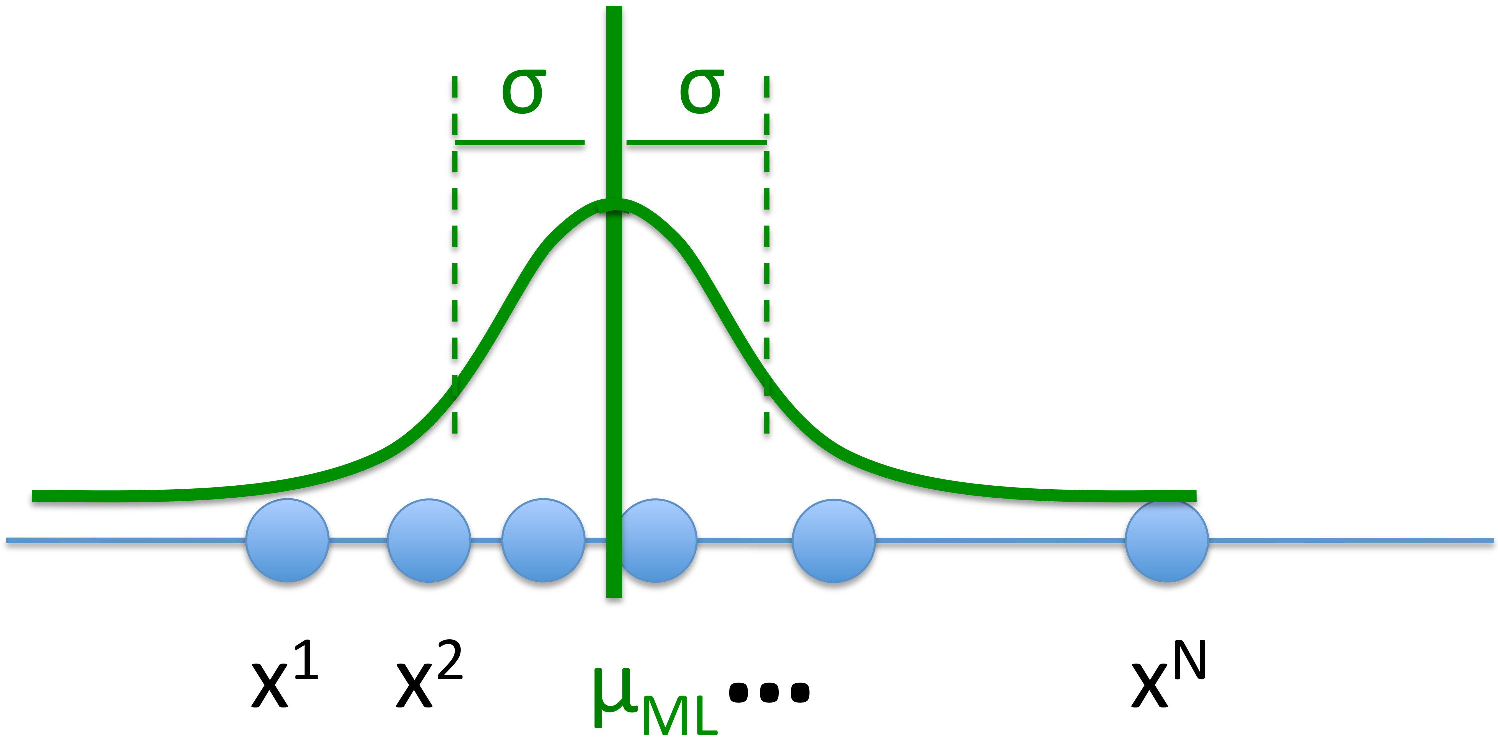
- We can write down the **likelihood function**:

$$p(D \mid \mu) = \prod_{i=1}^{N} p(x^i \mid \mu, \sigma) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma}} \exp\left\{ -\frac{1}{2\sigma^2}(x^i - \mu)^2 \right\}$$

- We want to choose the μ that maximizes this expression
  - Take log, then basic calculus: differentiate w.r.t. μ, set derivative to 0, solve for μ to get **sample mean**

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

Example: Maximum Likelihood For a 1D Gaussian



Maximum Likelihood

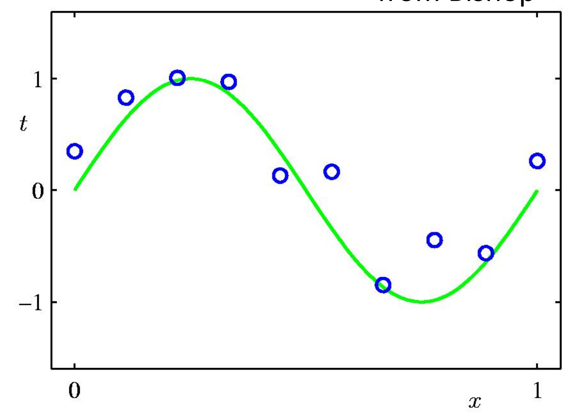
Least-squares Regression

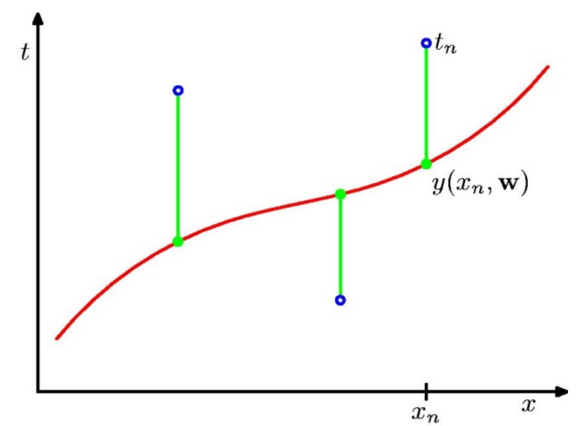
- Standard loss/cost/objective function measures the squared error in y [the prediction of t(x)] from x.

$$J(\mathbf{w}) = \sum_{n=1}^{N} [t^{(n)} - y^{(n)}]^2$$

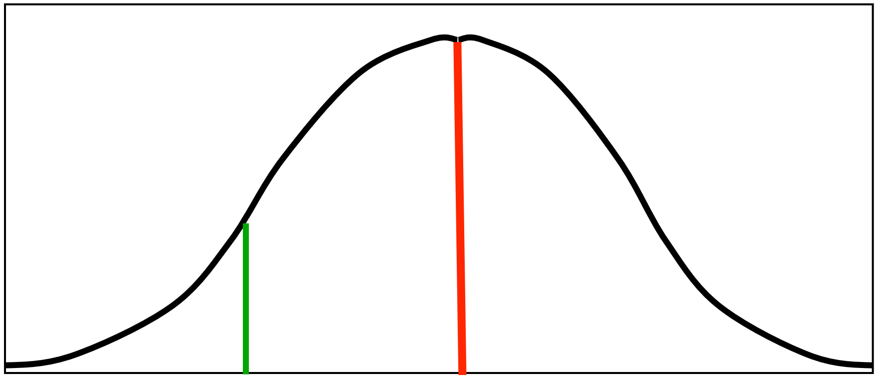
- The loss for the red hypothesis is the sum of the squared vertical errors.

from Bishop





When is minimizing the squared error equivalent to Maximum Likelihood Learning?



t = the correct answer

y = model's estimate of most probable value

Minimizing the squared residuals is equivalent to maximizing the log probability of the correct answer under a Gaussian centered at the model's guess.

$$y^{(n)} = y(\mathbf{x}^{(n)}, \mathbf{w})$$

$$p(t^{(n)} \mid y^{(n)}) = p(y^{(n)} + noise = t^{(n)} \mid \mathbf{x}^{(n)}, \mathbf{w}) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(t^{(n)} - y^{(n)})^2}{2\sigma^2}}$$

$$-\log p(t^{(n)} \mid y^{(n)}) = \log\sqrt{2\pi} + \log\sigma + \frac{(t^{(n)} - y^{(n)})^2}{2\sigma^2}$$

can be ignored if sigma is fixed

can be ignored if sigma is same for every case

Minimizing the absolute error

- An alternative to the least-squares objective:

$$\min_{over\ \mathbf{w}} \quad \sum_n | t^{(n)} - \mathbf{w}^T \mathbf{x}^{(n)} |$$

- This minimization involves solving a linear programming problem.
- It corresponds to maximum likelihood estimation if the output noise is modeled by a Laplacian instead of a Gaussian.

$$p(t^{(n)} \mid y^{(n)}) \;=\; a\, e^{-a |t^{(n)} - y^{(n)}|}$$

$$-\log p(t^{(n)} \mid y^{(n)}) \;=\; -a \,| t^{(n)} - y^{(n)} | + const$$

Regularized least squares

Increasing the input features this way can complicate the model considerably

Aim: select the appropriate model complexity automatically

Standard approach: regularization

$$\tilde{J}(\mathbf{w}) = \sum_{n=1}^{N} \{y(\mathbf{x}^{(n)}, \mathbf{w}) - t^{(n)}\}^2 \ + \ \alpha \mathbf{w}^T \mathbf{w}$$

The penalty on the squared weights is known as ridge regression in statistics

Leads to modified update rule

$$w \leftarrow w + 2\lambda[(t^{(n)} - y(x^{(n)}))x^{(n)} - \alpha w]$$