Roadmap, Jan 20

SRTE Quote:

The course kind of felt like it was in disarray most of the time, but it worked out in the end.

What We've Done so Far

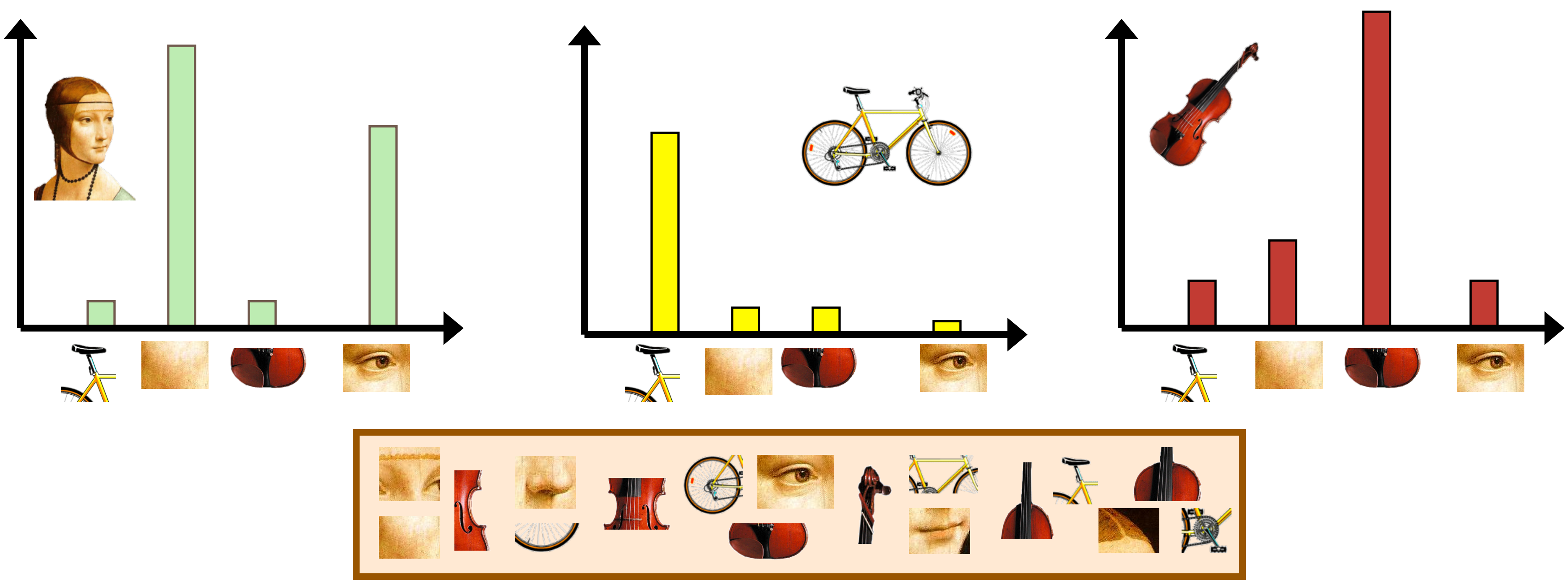
- Reading: Chaps 2-4 of Prince Book
- Review of probability theory
- MLE vs MAP
  - Frequentist vs Bayesian

Where we are Heading

- Bag of (Visual) Words Modeling
  - Scene recognition; Object recognition

Bag of Words Overview

1. Extract features — sparse or dense points in continuous vector space
2. Learn "visual vocabulary" — Training set. Clustering to get visual "words" e.g. K-means
3. Quantize features using visual vocabulary
4. Represent images by frequencies of "visual words" — MAP estimate of categorical distribution



Where we are Heading

- MLE/MAP estimation of Categorical Distribution
  - Today: revisit Bernoulli with new notation that more easily generalizes to Categorical (Homework 1)
- Feature Extraction (Read 13.1-13.3 of Prince)
- Clustering
  - K-means derivation (most common)
  - Other alternatives: mean-shift, K-medoids, quickshift
- Bag of Words (Read 20.1-20.2 of Prince)

- Also, Readings and Critiques of BoW papers.

Where we are Heading

- MLE/MAP estimation of Categorical Distribution
  - Today: revisit Bernoulli with new notation that more easily generalizes to Categorical **(Homework 1)**
- Feature Extraction **(Read 13.1-13.3 of Prince)**
- Clustering
  - K-means derivation (most common)
  - Other alternatives: mean-shift, K-medoids, quickshift
- Bag of Words **(Read 20.1-20.2 of Prince)**

- Also, **Readings and Critiques of BoW papers**.

Conjugate Distributions

We need probability distributions over model parameters as well as over data and world state. Hence, some distributions describe the parameters of others:

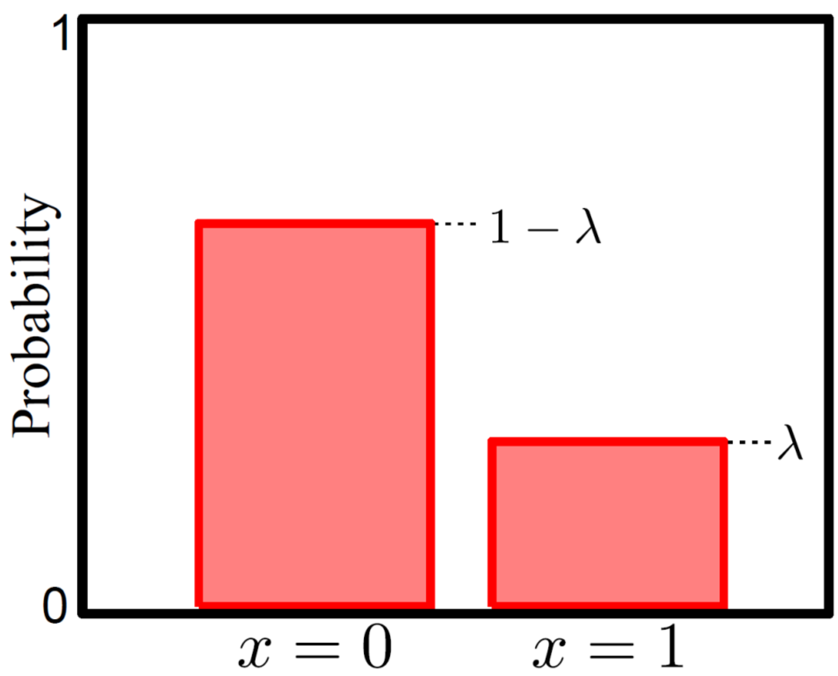
Distribution	**Domain**	**Parameters modeled by**
Bernoulli	$x \in \{0, 1\}$	beta
categorical	$x \in \{1, 2, \ldots, K\}$	Dirichlet
univariate normal	$x \in \mathbb{R}$	normal inverse gamma
multivariate normal	$\mathbf{x} \in \mathbb{R}^k$	normal inverse Wishart

Conjugate Distributions

We need probability distributions over model parameters as well as over data and world state. Hence, some distributions describe the parameters of others:

Distribution	**Domain**	**Parameters modeled by**
Bernoulli	$x \in \{0, 1\}$	beta
categorical	$x \in \{1, 2, \ldots, K\}$	Dirichlet
univariate normal	$x \in \mathbb{R}$	normal inverse gamma
multivariate normal	$\mathbf{x} \in \mathbb{R}^k$	normal inverse Wishart

Bernoulli Distribution



$$
\begin{aligned}
Pr(x=0) &= 1-\lambda \\
Pr(x=1) &= \lambda.
\end{aligned}
$$

or

$$Pr(x) = \lambda^x (1-\lambda)^{1-x}$$

For short we write:

$$\boxed{Pr(x) = \text{Bern}_x[\lambda]}$$

Bernoulli distribution describes situation where only two possible outcomes y=0/y=1 or failure/success

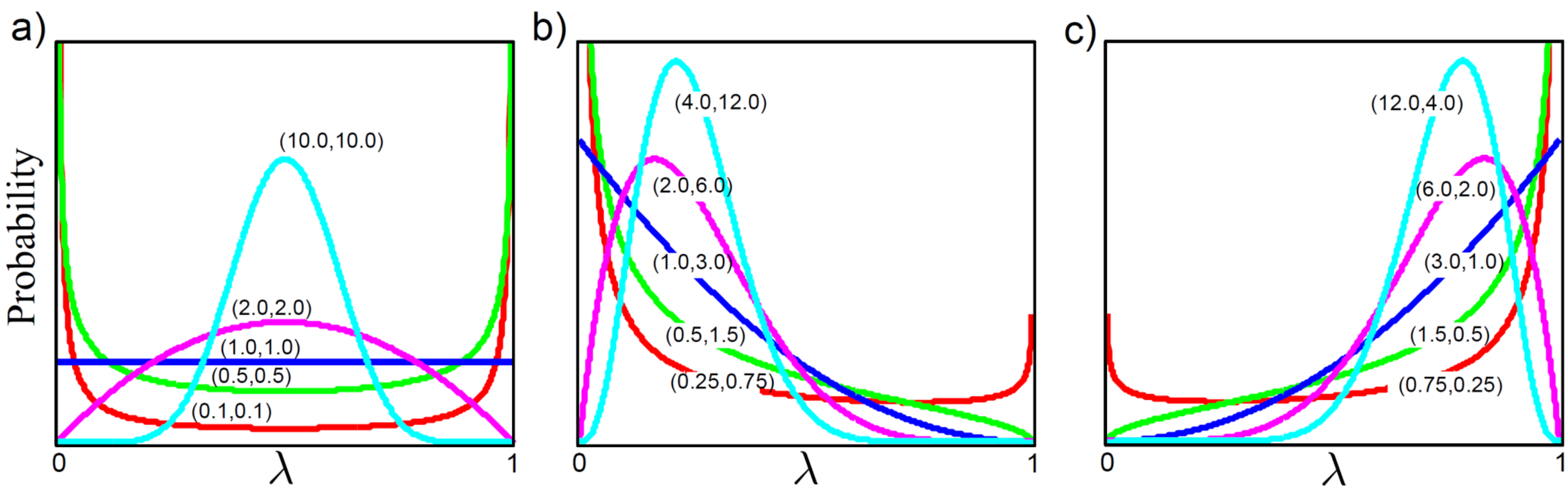
Takes a single parameter $\lambda \in [0,1]$

Beta Distribution

Defined over data $\lambda \in [0,1]$ (i.e. parameter of Bernoulli)

$$Pr(\lambda) = \frac{\Gamma[\alpha+\beta]}{\Gamma[\alpha]\Gamma[\beta]}\lambda^{\alpha-1}(1-\lambda)^{\beta-1}$$

$$\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$$
$$\Gamma(z) = (z-1)!$$

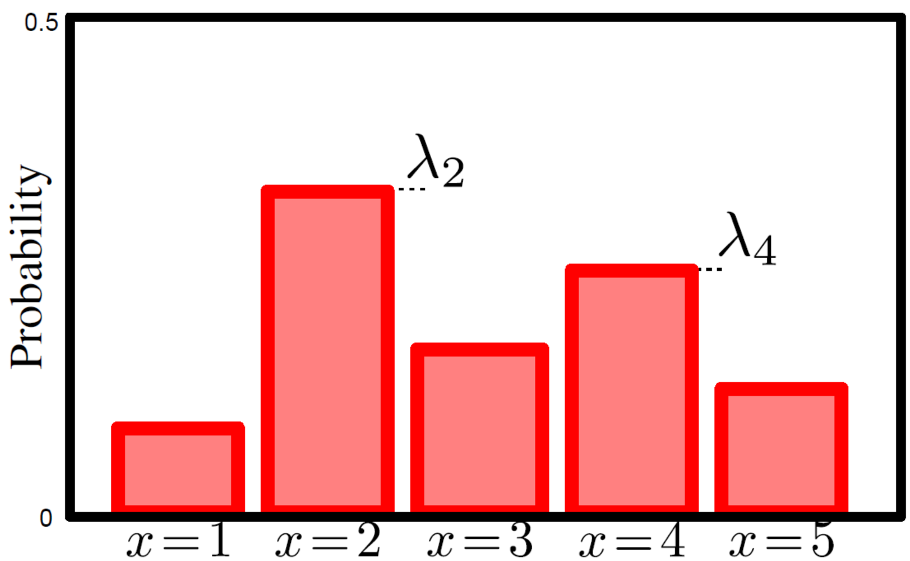


- Two parameters α,β both > 0
- Mean depends on relative values $E[\lambda] = \alpha/(\alpha+\beta)$.
- Concentration depends on magnitude

For short we write:

$$Pr(\lambda) = \text{Beta}_\lambda[\alpha,\beta]$$

Categorical Distribution



$$Pr(x=k)=\lambda_k$$

or can think of data as vector with all elements zero except k^th e.g. $\mathbf{e}_4$ = [0,0,0,1,0]

$$Pr(\mathbf{x}=\mathbf{e}_k)=\prod_{j=1}^{K}\lambda_j^{x_j}=\lambda_k$$

For short we write:

$$\boxed{Pr(x)=\text{Cat}_x\left[\boldsymbol{\lambda}\right]}$$

Categorical distribution describes situation where K possible outcomes y=1… y=k.

Takes K parameters $\lambda_k \in [0,1]$ where $\sum_k \lambda_k = 1$

Dirichlet Distribution

Defined over K values $\lambda_k \in [0,1]$ where $\sum_k \lambda_k = 1$

$$Pr(\lambda_1 \ldots \lambda_K) = \frac{\Gamma[\sum_{k=1}^K \alpha_k]}{\prod_{k=1}^K \Gamma[\alpha_k]} \prod_{k=1}^K \lambda_k^{\alpha_k - 1}$$

Has k parameters $\alpha_k > 0$

Or for short: $\boxed{Pr(\lambda_1 \ldots \lambda_K) = \mathrm{Dir}_{\lambda_{1\ldots K}}[\alpha_1, \alpha_2, \ldots, \alpha_K]}$