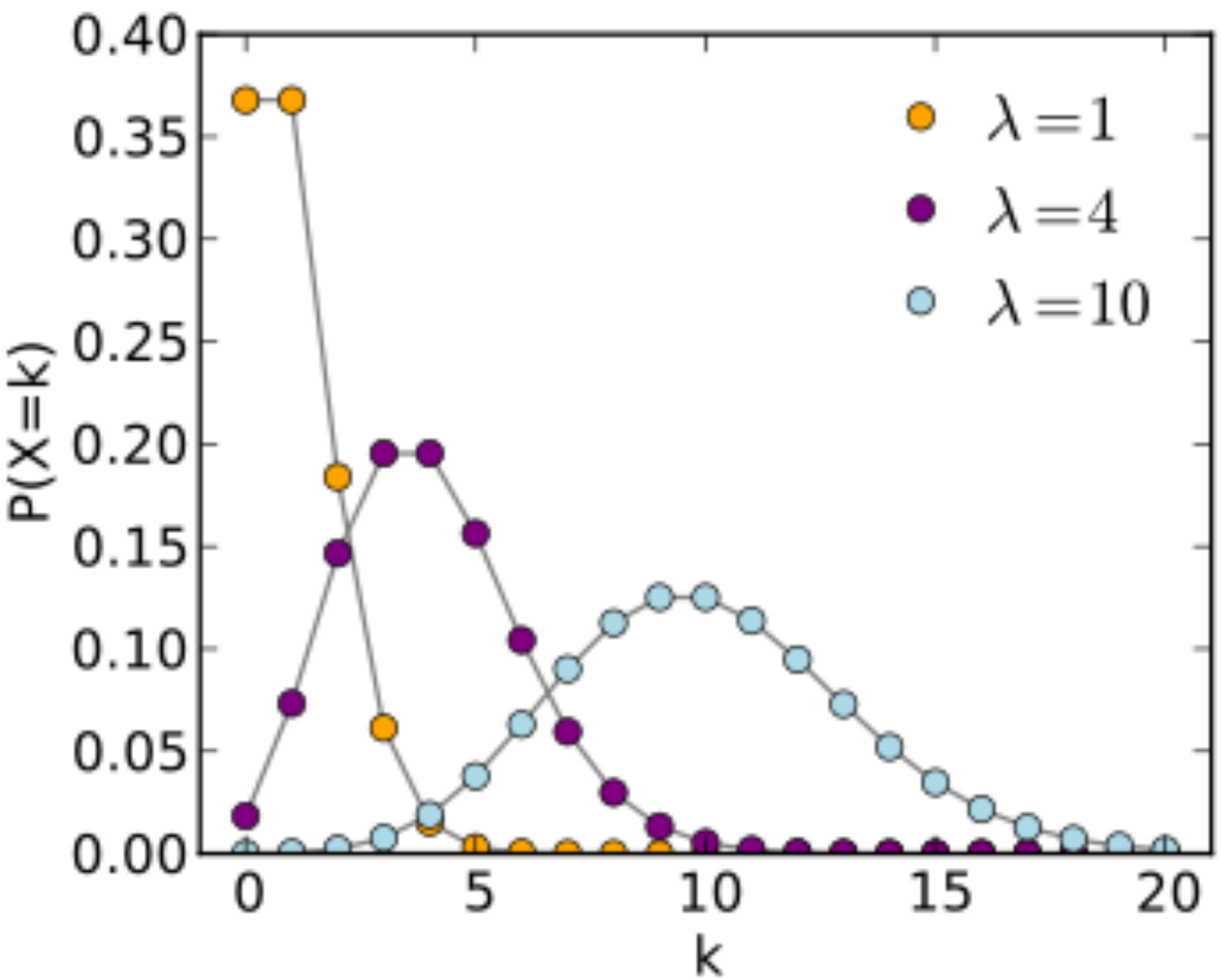
Generalized linear models



Comments (Oct. 10)

- Thought questions due this Thursday
- Office hours today shifted by 1 hour (starting at 4 p.m., ending at 5:30 p.m.)

Summary so far

- From chapters 1 and 2, obtained tools needed to talk about uncertainty/noise underlying machine learning
  - capture uncertainty about data/observations using probabilities
  - formalize estimation problem for distributions
- Identify variables x_1, …, x_d
  - e.g. observed features, observed targets
- Pick the desired distribution
  - e.g. p(x_1, …, x_d) or p(x_1 | x_2, …, x_d) (conditional distribution)
  - e.g. p(x_i) is Poisson or p(y | x_1, …, x_d) is Gaussian
- Perform parameter estimation for chosen distribution
  - e.g., estimate lambda for Poisson
  - e.g. estimate mu and sigma for Gaussian

Summary so far (2)

- For prediction problems, which is much of machine learning, first discuss
  - the types of data we get (i.e., features and types of targets)
  - goal to minimize expected cost of incorrect predictions
- Starting from this general problem specification, it is useful to use our parameter estimation techniques to solve this problem
  - e.g., specify Y = Xw + noise, estimate mu = xw
- Underlying assumptions
  - iid data, so log of likelihood splits up into sum
  - noise is independent of features

Summary so far (3)

- For linear regression setting, modeling p(y|x) as a Gaussian with mu = <x,w> and a constant sigma

- Performed maximum likelihood to get weights w

- Possible question: why all this machinery to get to linear regression?
  - one answer: makes our assumptions about uncertainty more clear
  - another answer: it will make it easier to generalize p(y | x) to other distributions (which we will do with GLMs)

Exercise: MAP for Poisson

- Recall we estimated lambda for Poisson p(x)
  - Had a dataset of scalars {x1, …, xn}
  - For MLE, found the closed form solution lambda = average of xi
- Can we use gradient descent for this optimization?

Exercise: Linear regression

- Recall we estimated w for p(y l x) as a Gaussian

- We discussed the closed form solution

$$\mathbf{w} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

- and using batch or stochastic gradient descent

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \mathbf{X}^\top (\mathbf{X}\mathbf{w}_t - \mathbf{y})$$

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \mathbf{x}_t^\top (\mathbf{x}_t \mathbf{w}_t - y_t)$$

- Now imagine you have 10 new data points. How do we get a new w, that incorporates these data points?

Exercise: Predicting the number of accidents

- In Assignment 1, learned p(y) as Poisson, where Y is the number of accidents in a factory
- How would the question from assignment 1 change if we also wanted to condition on features?
  - For example, want to model the number of accidents in the factory, given x1 = size of the factory and x2 = number of employees
- What is p(y I x)? What are the parameters?

Whiteboard

- Generalized linear models
  - Poisson regression
  - Logistic regression (intro)
  - General exponential family models

Exercise

- Why is ML and MAP estimation seemingly more complicated for regression setting than parameter estimation in third chapter?
  - e.g., previously estimated parameter lambda for Poisson p(x | lambda)
- For estimating p(y | x) as a Poisson distribution, we did not have a closed for solution for w, but we did for lambda when estimating Poisson p(x)
- Reason: conditional distribution lambda = exp<x , w>, rather than just directly estimating one lambda for p(x)