COMPUTER SCIENCE
INDIANA UNIVERSITY
School of Informatics and Computing
Bloomington

# Generalized linear models

# Reminders/Comments

- Thought questions due today
- Notations file updates
  - e.g., exp(-x) is e^{-x}
- Look at date in notes to know when updated

# 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)
  - the costs associated with incorrect predictions
  - specify the desire 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
  - potentially other assumptions (like noise is independent of features)

# Summary so far (3)

- Discussed notion of the "optimal" thing to do
  - this section on Bayes optimal was mostly to give intuition about how this might be formalized; if it is confusing you, it did not have its intended purpose and you can mostly ignore it as you will not be tested on it
- Optimal if we can specify individual values for each x
  - this would never be possible in practice, unless x is discrete and small
- Optimality if have a restricted class of functions
  - e.g., $\mathcal{F} = \{f : \mathbb{R}^d \rightarrow \mathbb{R} \mid f(\mathbf{x}) = \mathbf{x}^\top \mathbf{w} \text{ for } \mathbf{w} \in \mathbb{R}^d\}$
  - implicitly, we think of this set as $\mathcal{F} = \{\mathbf{w} \in \mathbb{R}^d \mid \mathbf{w} \in \mathbb{R}^d\}$
  - clearly for this set, cannot specify f(x) individual, as tied by w
- For now, we will not try to weight over all functions in F; we are going to pick a point estimate (which corresponds to MAP)

# Summary so far (4)

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

- Perform point-estimation (maximum likelihood and MAP) to get weights w (rather than weighting across multiple “good” w)

- For linear regression, parametrized mu = f(x) = <x, w>

- Possible question: why all this machinery to get to linear regression?
  - one answer: makes our assumptions about uncertainty more clear
  - another answer: gives us nice (convex) optimizations (we’ll see this now)

# **Example**: linear regression

- For the Gaussian distribution, why did we parametrize mu with f(x) = xw a linear function, in p(y l x)?

- What if we picked a different function? e.g., polynomial, sigmoid, neural network, or generally any non-linear function

- If we picked a constant function (like sine), no parameters to learn, expert has specified mu = sine(x)
  - this seems like a poor choice, lets not do this

- Otherwise, imagine f is some generic non-linear function parameterized by w (not necessarily linear)

- How do we get the MAP estimate?
  - using knowledge of models makes optimization specification easier

# **Example**: regularization and bias

- Remember that we picked a prior (Gaussian or Laplace) and obtained (l2 or l1) regularization

- We discussed the bias of this regularization
  - no regularization was unbiased E[w*] = true w
  - with regularization meant E[w*] was not equal to the true w

$$\mathbb{E}[(w^* - w_{\text{true}})^2] = \text{Bias}(w^*, w_{\text{true}})^2 + \text{Variance}(w^*)$$

- Previously, however, mentioned that MAP and ML converge to the same estimate

- Does that happen here?

# Whiteboard

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