Course Review

Reminders/comments

- Mini-Project due this Friday
- Final exam: 2:00-4:00 p.m., Friday, December 15
  - ETLC E1 013
  - Seating will be randomized, so you will need to find your seat
- Please complete the online course evaluation:
  - https://usri.srv.ualberta.ca/etw/ets/et.asp?nxappid=WCQ&nxmid=start

Probability review

- Quantify uncertainty using probability theory
- Discussed sigma-algebras and probability measures
- Discussed random variables as functions of event-space
- Discussed relationships between random variables, including (in)dependence and conditional independence
- Discussed operations, like expected value, marginalization, Bayes rule, chain rule

Exercise: probability

- Suppose that we have created a machine learning algorithm that predicts whether a link will be clicked with 99% sensitivity (TPR) and 99% specificity (FPR). The rate the link is actually clicked is 1/1000 visits to a website. If we predict the link will be clicked on a specific visit, what is the probability it will actually be clicked?

- Let C be binary RV, with C = 1 indicating predict click

- p(C = 1 | y = 1) = TPR

- p(C = 1 | y = 0) = 1-FPR

MAP and ML

- See a set of random variables X1, …, Xn
- We assumed that these RVs are independent and identically distributed
- Then we assumed a distribution for each p(Xi | theta)
  - e.g., p(Xi | theta) is a Gaussian
- How would our ML objectives change if we did not assume independence?
  - previously maximized p(D | theta) = prod_{i=1}^n p(Xi | theta)

Exercise: understand behaviour of algorithms

- Example: run NN twice on the same data. Should you expect to get about the same error on the testing set?

- Example: give NN the same training data, in the same order, with the same starting point —> should it produce the same final set of weights?

- Example: when performing batch gradient descent with line search, should you expect loss(wt) to consistently decrease?

- Example: when performing stochastic gradient descent with a small fixed stepsize, should you expect loss(wt) to consistently decrease?

Generalized linear models

- Generalize distribution p(y l x) to any exponential family model
- Result: learning parameters w such that

$$1.\ g\left(E[y|\mathbf{x}]\right) = \boldsymbol{\omega}^T\mathbf{x}$$

$$2.\ p(y|\mathbf{x}) \in \text{Exponential Family}$$

Generalized linear models

- Can pick any natural exponential family distribution for p(y | x)
- If p(y | x) is Gaussian, then we get linear regression with <x, w> approximating E[y | x]
- If p(y | x) is Bernoulli, then we get logistic regression with sigmoid(<x, w>) approximating E[y | x]
- If p(y | x) is Poisson, then we get Poisson regression with exp(<x, w>) approximating E[y | x]
- If p(y | x) is a Multinomial (multiclass), then we get multinomial logistic regression with softmax(<x, w>) approximating E[y | x]
- For all of these, just estimating w to get this dot product

Generalized linear models

- Generalize distribution p(y | x) to exponential family model
- Result: learning parameters w such that

$$1.\ g\left(E[y|\mathbf{x}]\right) = \boldsymbol{\omega}^T\mathbf{x}$$

$$2.\ p(y|\mathbf{x}) \in \text{Exponential Family}$$

- Is this a nonlinear model?

Classification

- Logistic regression
- Multinomial logistic regression
- Naive Bayes

Exercise

- What model might you use if
  - we have binary features and targets?
  - binary targets and continuous features?
  - positive targets?
  - categorical features with a large number of categories?
  - multi-class targets, with continuous features?
- When might logistic regression do better than linear regression?
- When might Poisson regression do better than linear regression?

Generative vs Discriminative

- Logistic regression: learned p(y l x)
- Naive Bayes: learned p(x l y) and p(y)
- When might you think naive Bayes might be better? When might logistic regression be better?
  - hard to say for sure, but hypothesize based on your understanding
  - e.g., which one might have higher/lower bias and higher/lower variance?

Exercise: How do we make naive Bayes and logistic regression more focused on recent samples?

- Imagine the data is slowly drifting
  - e.g., trends in human population
  - e.g., components of a physical system (robot) slowly wear out

- What incremental learning approach might make logistic regression more reactive to recent data?
  - e.g., batch approach versus stochastic approach

- What about updating our naive Bayes models?
  - recall the solutions are sample averages for each class, so it was straightforward to compute a running mean and running variance

Computing numerically stable running mean and variance

$$\mu_t = \frac{(t-1)}{t}\mu_{t-1} + \frac{x_t}{t}$$

$$= \mu_{t-1} + \frac{(x_t - \mu_{t-1})}{t}$$

$$s_t = s_{t-1} + \frac{(x_t^2 - s_{t-1})}{t}$$

$$\sigma_t^2 \leftarrow s_t - \mu_t^2$$

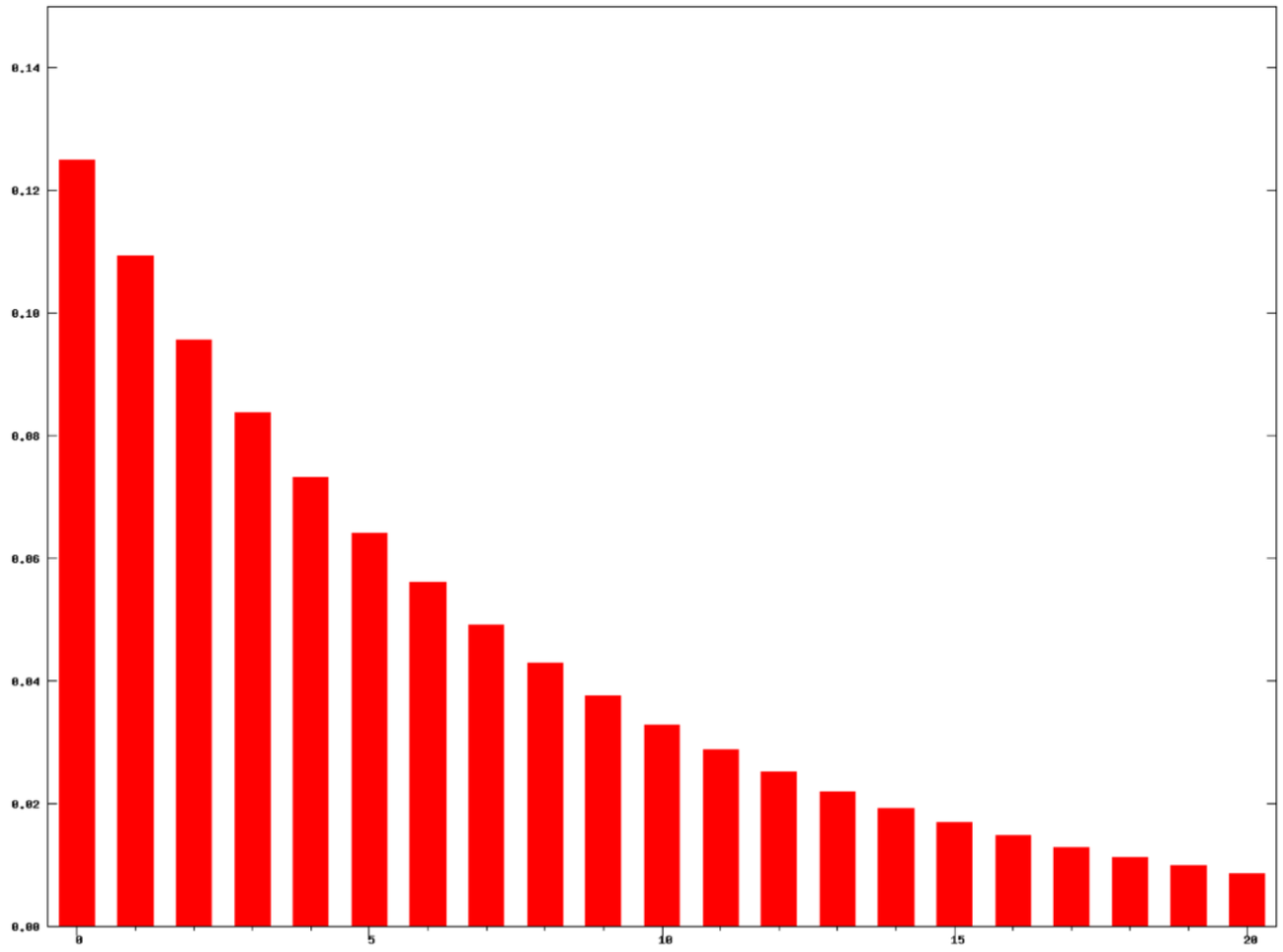
- Another algorithm: Welford's algorithm

Exponential average

$$\begin{aligned}\mu_t &= \alpha x_t + (1-\alpha)\mu_{t-1} \\ &= \alpha x_t + (1-\alpha)(\alpha x_{t-1} + (1-\alpha)\mu_{t-2}) \\ &= \alpha x_t + \alpha(1-\alpha)x_{t-1} + (1-\alpha)^2(\alpha x_{t-2} + (1-\alpha)\mu_{t-3}) \\ &= \alpha \sum_{i=0}^{\infty}(1-\alpha)^i x_{t-i}\end{aligned}$$

where $0 < \alpha < 1$

$$\sum_{i=0}^{\infty}(1-\alpha)^i = \frac{1}{\alpha}$$



Representation learning

- Fixed representations
  - polynomial expansions
  - Radial basis function networks — can be learning here, in terms of selecting centers or parameters to kernels
- Learned representations
  - neural networks
  - matrix factorization, for dictionary learning such as sparse coding

Whiteboard

- Practice final overview
- General comments:
  - In some cases lots of space, but doesn't mean you need to fill the entire page. If the answer can be concisely state in 2 sentences, that is perfectly reasonable
  - I am not looking for one specific "right" answer. I am looking for your thought process, to see if you understood the material
  - In the past, the wrong answer has been given (e.g., true, false question), but I gave full marks because the reasoning demonstrated understanding
  - Two common mistakes: (a) giving multiple answers, in case one is right. You'll lose marks for this (b) answering a different question than the one asked (read the question)