What about continuous variables?

- Billionaire says: If I am measuring a continuous variable, what can you do for me?
- You say: Let me tell you about Gaussians…

\[
P(x \mid \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}
\]
Some properties of Gaussians

- **Affine transformation (multiplying by scalar and adding a constant)**
  - \( X \sim N(\mu, \sigma^2) \)
  - \( Y = aX + b \rightarrow Y \sim N(a\mu + b, a^2\sigma^2) \)

- **Sum of Gaussians**
  - \( X \sim N(\mu_X, \sigma^2_X) \)
  - \( Y \sim N(\mu_Y, \sigma^2_Y) \)
  - \( Z = X + Y \rightarrow Z \sim N(\mu_X + \mu_Y, \sigma^2_X + \sigma^2_Y) \)

Learning a Gaussian

- **Collect a bunch of data**
  - Hopefully, i.i.d. samples
  - e.g., exam scores

- **Learn parameters**
  - Mean
  - Variance

\[
P(x \mid \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}
\]
MLE for Gaussian

- Prob. of i.i.d. samples $D=\{x_1,\ldots, x_N\}$:

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

- Log-likelihood of data:

$$\ln P(D \mid \mu, \sigma) = \ln \left( \left( \frac{1}{\sigma \sqrt{2\pi}} \right)^N \prod_{i=1}^{N} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \right)$$

$$= -N \ln \sigma \sqrt{2\pi} - \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2}$$

Your second learning algorithm:

MLE for mean of a Gaussian

- What’s MLE for mean?

$$\frac{d}{d\mu} \ln P(D \mid \mu, \sigma) = \frac{d}{d\mu} \left[ -N \ln \sigma \sqrt{2\pi} - \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2} \right]$$
MLE for variance

Again, set derivative to zero:

\[
\frac{d}{d\sigma} \ln P(D | \mu, \sigma) = \frac{d}{d\sigma} \left[ -N \ln \sigma \sqrt{2\pi} - \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2} \right] = \frac{d}{d\sigma} \left[ -N \ln \sigma \sqrt{2\pi} \right] - \sum_{i=1}^{N} \frac{d}{d\sigma} \left[ \frac{(x_i - \mu)^2}{2\sigma^2} \right]
\]

Learning Gaussian parameters

MLE:

\[
\hat{\mu}_{MLE} = \frac{1}{N} \sum_{i=1}^{N} x_i
\]

\[
\hat{\sigma}^2_{MLE} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{\mu})^2
\]

BTW. MLE for the variance of a Gaussian is **biased**

- Expected result of estimation is not true parameter!
- Unbiased variance estimator:

\[
\hat{\sigma}^2_{unbiased} = \frac{1}{N - 1} \sum_{i=1}^{N} (x_i - \hat{\mu})^2
\]
Bayesian learning of Gaussian parameters

- Conjugate priors
  - Mean: Gaussian prior
  - Variance: Wishart Distribution

Prior for mean:

\[
P(\mu | \eta, \lambda) = \frac{1}{\lambda \sqrt{2\pi}} e^{-\frac{(\mu - \eta)^2}{2\lambda^2}}
\]

MAP for mean of Gaussian

\[
P(\mu | \eta, \lambda) = \frac{1}{\lambda \sqrt{2\pi}} e^{-\frac{(\mu - \eta)^2}{2\lambda^2}}
\]

\[
P(D | \mu, \sigma) = \left(\frac{1}{\sigma \sqrt{2\pi}}\right)^N \prod_{i=1}^{N} e^{-\frac{(y_i - \mu)^2}{2\sigma^2}}
\]

\[
\frac{d}{d\mu} \left[ \ln P(D | \mu) P(\mu) \right] = \frac{d}{d\mu} \left[ \ln P(D | \mu) + \ln P(\mu) \right]
\]
Prediction of continuous variables

- Billionaire says: Wait, that’s not what I meant!
- You say: Chill out, dude.
- He says: I want to predict a continuous variable for continuous inputs: I want to predict salaries from GPA.
- You say: I can regress that…

The regression problem

- Instances: \(<x_j, t_j>\)
- Learn: Mapping from x to t(x)
- Hypothesis space:
  - Given, basis functions \( H = \{h_1, \ldots, h_K\} \)
  - Find coeffs \( w = (w_1, \ldots, w_k) \)
  \[ t(x) \approx \hat{f}(x) = \sum_i w_i h_i(x) \]
  - Why is this called linear regression???
    - model is linear in the parameters

- Precisely, minimize the residual squared error:
  \[
  w^* = \arg \min_w \sum_j \left( t(x_j) - \sum_i w_i h_i(x_j) \right)^2
  \]
The regression problem in matrix notation

\[ w^* = \arg \min_w \sum_j \left( t(x_j) - \sum_i w_i h_i(x_j) \right)^2 \]

\[ w^* = \arg \min_w \underbrace{(Hw - t)^T(Hw - t)}_{\text{residual error}} \]

Regression solution = simple matrix operations

solution: \[ w^* = \begin{bmatrix} \underbrace{H^T H}^{-1} & H^T \end{bmatrix} \begin{bmatrix} b \\ A^{-1} \end{bmatrix} = A^{-1} b \]

where \[ A = H^T H = \begin{bmatrix} k \times k \text{ matrix} & \text{for } k \text{ basis functions} \\ b = H^T t = \begin{bmatrix} b \\ k \times 1 \text{ vector} \end{bmatrix} \]
But, why?

- Billionaire (again) says: Why sum squared error???
- You say: Gaussians, Dr. Gateson, Gaussians...

- Model: prediction is linear function plus Gaussian noise
  \[ t = \sum_i w_i h_i(x) + \varepsilon \]

- Learn \( w \) using MLE
  \[
P(t \mid x, w, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{[t - \sum_i w_i h_i(x)]^2}{2\sigma^2}}
  \]

Maximizing log-likelihood

Maximize:

\[
\ln P(D \mid w, \sigma) = \ln \left( \frac{1}{\sigma \sqrt{2\pi}} \right)^N \prod_{j=1}^N e^{-\frac{[t_j - \sum_i w_i h_i(x_j)]^2}{2\sigma^2}}
\]

Least-squares Linear Regression is MLE for Gaussians!!!
Applications Corner 1

- Predict stock value over time from
  - past values
  - other relevant vars
    - e.g., weather, demands, etc.

Applications Corner 2

- Measure temperatures at some locations
- Predict temperatures throughout the environment

[Guestrin et al. ’04]
Applications Corner 3

- Predict when a sensor will fail
  - based several variables
    - age, chemical exposure, number of hours used,…

Announcements 1

- Readings associated with each class
  - See course website for specific sections, extra links, and further details
  - Visit the website frequently

- Recitations
  - Thursdays, 5:00-6:20 in Wean Hall 5409

- Special recitation on Matlab
  - Sept. 18 Tue. 4:30-5:50pm NSH 3002

- Carlos away on Monday Sept. 17th
  - Prof. Eric Xing will teach the lecture
Announcement 2

- First homework out later today!
  - Download from course website!
  - Start early!!! :)
  - Due Oct 3rd

- To expedite grading:
  - there are 4 questions
  - please hand in 4 stapled separate parts, one for each question

Bias-Variance tradeoff – Intuition

- Model too “simple” → does not fit the data well
  - A biased solution

- Model too complex → small changes to the data, solution changes a lot
  - A high-variance solution
(Squared) Bias of learner

- Given dataset $D$ with $m$ samples, learn function $h(x)$
- If you sample a different datasets, you will learn different $h(x)$
- **Expected hypothesis**: $E_D[h(x)]$

- **Bias**: difference between what you expect to learn and truth
  - Measures how well you expect to represent true solution
  - Decreases with more complex model
  $$bias^2 = \int_x \{ E_D[h(x)] - t(x) \}^2 p(x)dx$$

Variance of learner

- Given a dataset $D$ with $m$ samples, you learn function $h(x)$
- If you sample a different datasets, you will learn different $h(x)$
- **Variance**: difference between what you expect to learn and what you learn from a from a particular dataset
  - Measures how sensitive learner is to specific dataset
  - Decreases with simpler model
  $$\overline{h}(x) = E_D[h(x)]$$
  $$\text{variance} = \int E_D[(h(x) - \overline{h}(x))^2]p(x)dx$$
Bias–Variance Tradeoff

- Choice of hypothesis class introduces learning bias
  - More complex class → less bias
  - More complex class → more variance

Bias–Variance decomposition of error

- Consider simple regression problem $f: X \rightarrow T$
  $$t = f(x) = g(x) + \epsilon$$

  - noise $\sim N(0, \sigma)$
  - deterministic

Collect some data, and learn a function $h(x)$
What are sources of prediction error?
Sources of error 1 – noise

- What if we have perfect learner, infinite data?
  - If our learning solution \( h(x) \) satisfies \( h(x) = g(x) \)
  - Still have remaining, *unavoidable error* of \( \sigma^2 \) due to noise \( \varepsilon \)

\[
error(h) = \int_x \int_t (h(x) - t)^2 p(f(x) = t|x) p(x) dt dx
\]

Sources of error 2 – Finite data

- What if we have imperfect learner, or only \( m \) training examples?
- What is our expected squared error per example?
  - Expectation taken over random training sets \( D \) of size \( m \), drawn from distribution \( P(X,T) \)

\[
E_D \left[ \int_x \int_t (h(x) - t)^2 p(f(x) = t|x) p(x) dt dx \right]
\]
Bias-Variance Decomposition of Error

Assume target function: \( t = f(x) = g(x) + \varepsilon \)

Then expected square error over fixed size training sets \( D \) drawn from \( P(X,T) \) can be expressed as sum of three components:

\[
E_D \left[ \int_x \int_t (h(x) - t)^2 p(t|x)p(x) dt dx \right] = \text{unavoidable Error} + \text{bias}^2 + \text{variance}
\]

Where:

\[
\text{unavoidable Error} = \sigma^2 \]

\[
\text{bias}^2 = \int (E_D[h(x)] - g(x))^2 p(x) dx
\]

\[
\bar{h}(x) = E_D[h(x)]
\]

\[
\text{variance} = \int E_D[(h(x) - \bar{h}(x))^2] p(x) dx
\]

What you need to know

- Gaussian estimation
  - MLE
  - Bayesian learning
  - MAP
- Regression
  - Basis function = features
  - Optimizing sum squared error
  - Relationship between regression and Gaussians
- Bias-Variance trade-off
- Play with Applet