CheatSheet Basics Machine Learning

Chu Duc Thang

April 2021

1 Chapter 1

Introduction to the class. Only 1 page, no important information here

2 Chapter 2: Probability

1. Sample space/outcome space vs event space:
   - Sample space: Ω
   - Event space: Subset of sample space, ex: powerset (discrete), Borel Field (continuous)
2. Discrete vs Continuous RV
   - Discrete: {}, N, words
   - Continuous: [], R, R^k
3. Probability mass function (pmf) vs probability density function (pdf)
   - Pmf: $\Omega \rightarrow [0, 1]$
   - Pdf: $\Omega \rightarrow [0, \infty)$, no singleton event, can be > 1
4. Special Distribution
   - Discrete: Uniform (n - #outcomes), Poisson (α - histogram/likely), Bernoulli (p - success)
   - Continuous: Gamma (α, β), Uniform (a, b), Normal(μ, σ), Exponential (α)
5. Marginal vs Conditional Distribution
   - Marginal: p(x) = $\sum_{y \in Y} p(x, y)$
   - Conditional: p(x‖y) = $\frac{p(y\|x)p(x)}{p(y)}$ or p(x,y,z) = p(x‖y,z)p(y‖z)p(z)
6. Expected value vs Conditional Expected value vs Variance

- E = $\sum_{x \in X} xp(x)$
- E[X$\|$Y] = $\sum_{x \in X} xp(x\|y)$
- Var = E[$(X - E[X])^2$] or $E[X^2] - E[X]^2$
- Properties of E: E[c] = c, E[cX] = cE[X], E[X + Y] = E[X] + E[Y], E[XY] = E[X]E[Y] (independence), E[E[Y$\|$X]] = E[Y]
- Properties of Var: Var[c] = 0, Var[cX] = c^2Var[X], Var[X + Y] = Var[X] + Var[Y] + 2Cov(X,Y)

7. Covariance vs Correlation
   - Cov = E[XY] - E[X]E[Y]
   - Corr = $\frac{Cov(x,y}{\sqrt{Var(x)}\sqrt{Var(y)}}$
   - Note: $-1 \leq Corr \leq 1$, but Cov is unbounded
8. Independence vs Conditional Independence
   - P(X,Y) = P(X)P(Y)
   - P(X,Y$\|$Z) = P(X$\|$Z)P(Y$\|$Z)

3 Chapter 3: Estimator

1. Formula $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
2. Bias: $E[\bar{X}] - E[X]$
3. Confidence interval: Pr(| $\bar{X} - E[\bar{X}]$ |$< \epsilon$) $>$ 1 - δ
   - $\mu \in [\bar{X} - \epsilon, \bar{X} + \epsilon]$
4. Chebyshev: Known variance and $\delta = \frac{\sigma^2}{n\epsilon^2}$
5. Hoeffding: Bounded between a and b
6. Convergence rate: How quickly the error has been reduced
7. Sample complexity:
   - As small as possible (data efficiency)
   - n $\geq \frac{v^2}{\delta\epsilon^2}$
8. Consistency: As n $\rightarrow \infty$, $\epsilon \rightarrow 0$ or $\bar{X} \rightarrow \mu$
   - Unbiased $\rightarrow$ consistency, but not the vice versa
9. Mean-squared error: MSE = Var(X) + $Bias(X)^2$

4 Chapter 4: Optimization

1. $w^* = argmin_w c(w)$
2. Closed form:
   - Stationary point (c'(w) = 0): local min, local max, saddle point
   - Global min: Boundary point or local min
   - Concave up vs Concave down: c"(w) > 0 → minimum vs c"(w) < 0 → maximum
   - Practical: non-convex function → not able to take derivative
3. Gradient Descent
   - Taylor series degree 2: Approximate the actual function, then taking the derivative of the approximated function
   - $w_{t+1} = w_t - \frac{c'(w_t)}{c''(w_t)}$
   - Difficult to compute $c''(w_t)$, constant stepsize η
   - Chossing stepsize: Too large (overshoot) vs too small (too long to converge)
   - Adaptive stepsize: $\eta_t = argmin_\eta c(w_t - \eta_t \nabla c(w_t))$
4. Properties of Optimization
   - argmin c(w) = argmax -c(w)
   - argmin c(w) = argmin ac(w) = argmin (c(w) ± a)
   - convex function

5 Chapter 5: MAP/MLE/Bayesian

1. p(w‖D) = $\frac{p(D\|w)p(w)}{p(D)}$
   - D: Data, w: parameter
   - Posterior = $\frac{\text{Likelihood}\times\text{Prior}}{\text{Evidence}}$
   - Point estimation (MLE/MAP) vs distribution estimation (Bayesian)
2. Maximum likelihood Estimation (MLE)
   - $argmax_w p(D\|w)$
   - Observed data is most probable
3. Maximum a posterior (MAP)
   - $argmax_w p(D\|w)p(w)$
   - Mode of the distribution

- As data increases, diminishes the prior knowledge
- If prior is good, MSE(w_{MAP}) < MSE(w_{MLE})
- Converge to w_{true} as long as p(w) $\neq$ 0 (prior)

4. Bayesian estimation
   - $\frac{p(D\|w)p(w)}{p(D)} = \frac{p(D\|w)p(w)}{\int p(D\|w)p(w)}$
   - Conjugate prior
   - Estimation of whole distribution
   - Taken account skewed, multi modal

6 Chapter 6: Optimal predictor

1. Setting: Passive/ Complete/ IID
2. Feature vs Target: Easy to gather vs Expensive to collect
3. Regression:
   - Setting: Y $\in$ R, [0, ∞]
   - Optimal cost: E[C] = $\int_X \int_Y cost(f(x), y)p(x, y)dydx$
   - Mean square: $f^*(x) = E[Y\|X]$
   - Absolute: $f^*(x) = Median[Y\|X]$
4. Classification:
   - Setting: Discrete, words, N, multi-label (1 input = $\geq$ 1 output), multi-class (1 input = 1 output), order vs no order
   - Optimal Cost: E[C] = $\int_X \sum_Y cost(f(x), y)p(x, y)dx$
   - $f^*(x) = argmaxp(y\|x)$
   - Minimize cost for each x and cannot have 0 cost
5. Irreducible vs Reducible error
   (a) Optimal E[C] = $\int_X p(x)Var[Y\|X = x]dx$ (irreducible = noise/variability in Y)
   (b) Sub-optimal: E[C] = $E[(f(x) - f^*(x))^2] + E[(f^*(x) - y)^2]$ (reducible + irreducible)
   (c) Reducible: better function $f^*(x)$
   (d) Irreducible: More features, but for the given dataset, cannot further reduce

7 Chapter 7: Linear/Polynomial Regression

1. $\epsilon \in N(0, \sigma^2)$
2. MLE: $w_{MLE} = \text{argmin} \sum_{i=1}^{n} (y_i - x_i^T w)^2$
3. Cumulative vs average error
4. Closed form formula (not invertible) vs GD (long time for large dataset)
5. Stochastic vs Mini-batch GD:
   - Unbiased estimation of gradient and stepsize decreases overtime
   - SGD is better for streaming data/online learning
6. Polynomial Regression
   - Fit of degree p
   - Benefit: Higher p $\rightarrow$ more expressive $\rightarrow$ reduce squared error
   - Drawback: Overfitting
   - p deg d features = $\binom{d+p}{p}$

8 Chapter 8: Generalization Error

1. Generalization vs Empirical
   - Generalization: Error on the test set
   - Empirical: Error on the training test
2. Overfitting
   - Complexity of model
   - Small dataset
   - Increase significantly in magnitude of coefficient
   - Mismatch training and testing error
3. Underfitting
   - Insufficient to represent true model
   - Increases the dataset
   - Increases the complexity of model
   - Training error high
4. Comparison of model:
   - Increases polynomial degree and plot training/test error

5. Hold-out test set
    - Disadvantage: Small dataset and used only once
    - Solution: K-fold and bootstrap sampling
6. Confidence Interval
    - Gaussian: tighter confidence, but usually unknown true variance
    - Student t's distribution: $S_m = \frac{1}{m}\sum_{i=1}^{m}(c_i(f) - \bar{X})^2$ and $\epsilon = t_{\delta,m-1}\frac{S_m}{\sqrt{m}}$
    - Not overlap $\rightarrow$ different
    - Low power-test
7. Parametric Test
    - p-value: Likelihood of observing outcomes if hypothesis is True
    - α: Smaller α, stronger evidence against H_o
    - Binomial Test: Compared 0,1 values
    - Paired t-test: Real value of error
8. Error
    - Type I: Reject H_o when it should not be rejected (wrong assumption)
    - Type II: Not reject H_o when it should be rejected (wide CI)

9 Chapter 9: Regularization

1. MAP: $w_j - N(0, \frac{\sigma^2}{\alpha})$
2. Gaussian norm
    - c(w) $= \frac{1}{2n}\sum_{i=1}^{n}(x_i^T w - y_i)^2 + \frac{\alpha}{2}\sum_{i=1}^{d} w_j^2$
3. Laplace norm
    - Not closed form $\rightarrow$ GD
    - Not differential at 0
4. Do not regularize w_0 since intercept term do not increase complexity or cause overfitting
5. Bias-Variance trade-off
    - Bias: Small bias $\rightarrow$ simpler function + regularization $\rightarrow$ small variance, but may underfitting
    - Variance: High variance $\rightarrow$ Complex function, but may overfitting

10 Chapter 10: Classification

1. Linear classifier
   - Non-linear function, equation of hyperplane
   - Reason: p(y = 1$\|$x,w) $\geq \alpha$ when $x^T w \geq 0$
   - Intercept term: Otherwise always go through origin
2. Conditional Bernoulli distribution: Output probability not actual label
3. α increases $\rightarrow$ more confidence
4. Consistently producing same result $\rightarrow$ good accuracy
5. Sigmoid function + Cross-entropy loss
6. GD: as usual
7. MSE: non-convex function $\rightarrow$ not guarantee to converge to a global minima